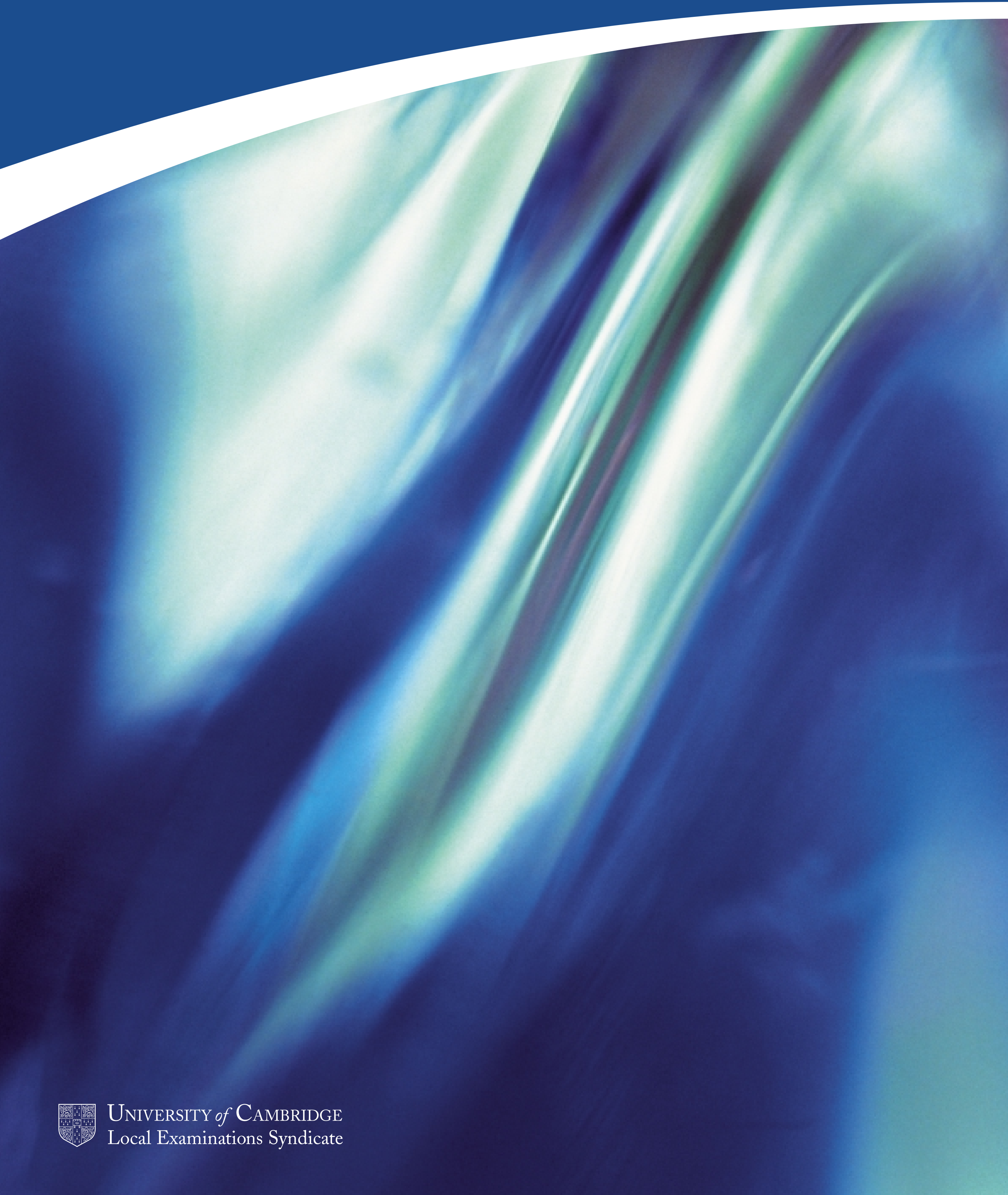
Issue 3 January 2007

Research Matters

CAMBRIDGE ASSESSMENT



UNIVERSITY *of* CAMBRIDGE
Local Examinations Syndicate

Citation

Articles in this publication should be cited as: Crisp, V. and Johnson, M. (2006). Examiners' annotations: Practice and purpose. *Research Matters: A Cambridge Assessment Publication*, 2, 10-12.

Research Matters : 3

A CAMBRIDGE ASSESSMENT PUBLICATION

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

The full issue and previous issues are available on our website
www.cambridgeassessment.org.uk/research

Foreword

I am pleased to introduce the third issue of *Research Matters*, which again seeks to stimulate debate and information exchange on matters central to assessment. Although many of the issues and lines of work described here will subsequently appear in articles in refereed journals, *Research Matters* provides a means of bringing them together into a single volume, allows early sight of key findings, and provides updates on developments germane to the assessment community. Indeed, *Research Matters* has begun to have an impact in its own right. The special issue on *Aspects of Writing* raised the profile of that work and led the government to commission an extension of the research. This reinforces the importance of dissemination. It is one thing to get the research done, but that is only half the task; genuine impact only comes through effective dissemination and the debates and exchanges which go with it. And in line with those who suggest that genuine change only comes when people begin to feel dissonance – feeling uncomfortable with the way things are – this edition of *Research Matters* does not seek to avoid controversy. John Rust's exploration of the application of the principles of psychometrics throws into relief the unhelpful nature of tribalism within assessment and measurement. Principles of measurement are fundamental to assessment, and false oppositions within the assessment community only impede development. I hope you find the items which drill down into marking and assessment of interest, and that the items which spark controversy and reflection open up lines of communication between different members of the assessment community.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

In this issue we report on topics ranging from the construct of Critical Thinking to the factors affecting examination success at A-level. In the opening article, Beth Black considers some of the literature on the definitions of Critical Thinking from philosophical and psychological perspectives. This is followed by two articles related to admissions tests. In the first of these John Bell considers the complexities involved in evaluating the predictive validity of selection tests. In the second article Joanne Emery and John Bell continue a discussion from *Research Matters*, Issue 1, on the difficulties of assessing high attaining candidates. This discussion takes place in the context of Thinking Skills Assessment.

The next two articles focus on A-level examinations and in the first of these Carmen Vidal Rodeiro and John Bell discuss factors that affect success at A-level based on information from different databases. This research was presented at the annual conference of the British Educational Research Association in September. The second article on A-levels, by John Bell, Eva Malacova, Carmen Vidal Rodeiro and Mark Shannon, discusses the claim that students are opting for allegedly easier subjects at A-level.

In his article on psychometric principles Professor John Rust outlines the fundamental principles by which the quality of assessments are judged and in his second article he describes the work of the Psychometrics Centre at Cambridge Assessment.

Martin Johnson's article considers the question of grading in competence-based qualifications in the light of recent national and international moves towards developing unified frameworks for linking qualifications. This article is a summary of an extended paper on grading issues currently in submission to the *Journal of Further and Higher Education*. We finish with 'Research News' which includes details of the new Statistical Reports Series as well as conference and seminar information.

Sylvia Green *Director of Research*

Critical Thinking – a tangible construct?

Beth Black Research Division

> *Are some outcomes of education too intangible to be measured? No doubt, there are some that we speak of often, like critical thinking.., that [is] so difficult to define satisfactorily that we have given up trying to define [it] specifically. To this extent, they are intangible [and] hard to measure.* (Ebel, 1965)

Forty years on from Ebel's quote, the testing of Critical Thinking has become a flourishing area. In the UK, tests which incorporate a Critical Thinking element include the BioMedical Admissions Test (BMAT), Thinking Skills Assessment (TSA), UniTest, UK Clinical Schools Admissions Test (UKCAT) and Watson Glaser Critical Thinking Appraisal UK (WGCTA-UK). Frequently the stated purpose of these tests is to help Higher Education establishments make admissions decisions, a situation with much precedent in the US where the Law Schools Admissions Test (LSAT) and Medical Colleges Admissions Test (MCAT) are de rigueur for applicants. It seems that to think critically is considered an advantageous or even essential ability for university students on some courses.

But what is Critical Thinking? Is Ebel's pessimistic view now outdated? This article hopes to introduce some of the debates within the construct of Critical Thinking and some of the implications for assessment of Critical Thinking. There are a number of protagonists within the field, and their definitions of what constitutes the construct of Critical Thinking vary enormously: 'chaos at the core' as Benderson wrote in 1990.

The early work of Robert H. Ennis, University of Illinois, propounded a 'pure skills' approach to Critical Thinking. Critical Thinking was defined as 'the correct assessing of statements' (Ennis quoted in Siegel, 1988) and was appended by a list of aspects of statement assessment and criteria. The caveat to this long list is that a complete set of criteria for Critical Thinking cannot be established, that 'intelligent judgement' is also required.

Thus, there are no clear boundaries defining the outer limits of what constitutes Critical Thinking. The implication of Ennis' early position (the 'pure skills' approach), is that if you can pass a test in Critical Thinking, you have Critical Thinking skills. The weakness in this definition is that someone may possess such skills and yet never use them. To *be* a critical thinker and not just be *able* to be one should be an important aspect of the definition. Ennis' (1996) later definition, 'Critical Thinking is reasonable, reflective thinking that is focused on deciding what to believe or do', introduces decision-making into the concept and the idea that Critical Thinking should affect a critical thinker's behaviour, that is, Critical Thinking is exercised and is not just pure skills.

Alec Fisher, Director of the Centre for Research in Critical Thinking at the University of East Anglia, insists that it must be a taught skill, and one that is transferable to other subject domains. He claims an important aspect is metacognition, that is, thinking about one's thinking. Arguably, metacognition can only be achieved through some conscious effort by reference to a good model of thinking. This is where the *teaching* of Critical Thinking comes into play. Additionally, Fisher argues that a critical thinker should exercise and apply these Critical Thinking skills not just in academic studies but in many situations (where appropriate). His definition is:

> *Critical Thinking is skilled and active interpretation and evaluation of observations and communications, information and argumentation.* (Fisher and Scriven, 1997)

Richard Paul, founder and director of Sonoma State University's Centre for Critical Thinking, argues that Critical Thinking courses often teach 'weak-sense' Critical Thinking, where the concepts within can become so atomistic that they are no longer Critical Thinking (just a series of 'moves'). Paul (1992) advocates Critical Thinking in a 'strong' sense. Critical thinkers should look at 'argument networks' or 'world views' and not merely reject an argument network on the basis of an atomistic flaw. One's deepest beliefs and ethical, moral and socio-cultural standpoints should be subject to Critical Thinking. Thus in order to think critically, one must use these skills on oneself; it is a reflective process.

> *Critical Thinking is disciplined, self-directed thinking which exemplifies the perfections of thinking appropriate to a particular mode or domain of thinking.*

John McPeck (1981) of the University of Western Ontario suggests that it cannot be taught as a standalone subject – one is always thinking about something – so that in theory one might offer Critical Thinking for Physics, or Critical Thinking for Geography.

> *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. [Critical Thinking] … is both conceptually and practically empty.*

In short, the construct of Critical Thinking is not precisely defined, nor is it the case that there is a single agreed definition.

Some of this division stems from the experts' fields (though all of the above are involved with the informal logic movement). Those from a philosophical background are interested in employing the tools of logic and reasoning in order to illuminate fundamental truths (with a tradition of more than 2,000 years of reasoning and argumentation). Meanwhile, those from a psychological background, for example, Sternberg and Halpern, are concerned with the thinking process and problem solving rather than logical reasoning. This tradition has evolved not from philosophical argument and discourse, but through experimentation on real subjects. Thus, psychologists may view the philosophers as giving an account of some 'ideal' Critical Thinking abilities, rather than actual performance where limiting factors (e.g. time, information, working memory capacity, motivation) come into play. There are differences between rules of logic and rules of thought. So, psychologists have been concerned with characterising Critical Thinking as it is performed under the limitations of the person and the context or environment. This notion is reflected in the definition of Professor Robert Sternberg (1986) of Yale University:

> *Critical Thinking comprises the mental processes, strategies, and representations people use to solve problems, make decisions, and learn new concepts.*

Thus, one expects from psychologists a more *descriptive* account of Critical Thinking, rather than an *aspirational* account.

Psychologists' definitions and taxonomies of Critical Thinking tend to emphasise problem solving rather than logic. Sternberg's psychological taxonomy of Critical Thinking skills involves metacomponents (e.g. formulating a strategy, monitoring progress in solving a problem), performance components (e.g. inductive and deductive reasoning, spatial visualisation) and knowledge-acquisition components (e.g. encoding and organising information). Interestingly, Critical Thinking tests which stem out of the cognitive tradition do not always separate out Critical Thinking from intelligence (e.g. Sternberg's Triarchic Test of Intellectual Skills).

Unsurprisingly, representatives from each tradition counter attack. Paul (quoted in Benderson) rejects the psychological account on the basis that the puzzles posed by psychologists as critical thinking teaching aids are self-contained or 'monological', that is, are simplistic in that they have a single correct answer and involve adopting just one frame of reference ('weak sense' Critical Thinking). 'True' Critical Thinking should involve 'multilogical' problems, involving multiple frames of reference or argument networks with no single correct answer; only then can a student reflect upon and evaluate their own beliefs. However, Sigel, an ETS researcher notes that 'Philosophers tend not to be empiricists… they just use themselves as sources of authorities. The psychologist is an empiricist who wants to create data that educators can then validate with their own experience.' (quoted in Benderson 1990)

Is there any definition to which the majority of experts would subscribe? Possibly the definition derived from a Delphi study[1] conducted in the United States by Facione (1990). In this study, 46 Critical Thinking experts, consisting of 24 panellists associated with philosophy (including Ennis and Paul), 9 associated with the social sciences, 2 with physical sciences and 10 with education formed a consensus on many aspects of Critical Thinking, including a definition and list of critical skills. The definition, quoted in full, reads as follows:

> *We understand Critical Thinking to be purposeful, self-regulatory judgement which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgement is based. CT [sic] is essential as a tool of inquiry. As such, CT is a liberating force in education and a powerful resource in one's personal and civic life. While not synonymous with good thinking, CT is a pervasive and self-rectifying human phenomenon. The ideal critical thinker is habitually inquisitive, well-informed, trustful of reason, open-minded, flexible, fair-minded in evaluation, honest in facing personal biases, prudent in making judgements, willing to reconsider, clear about issues, orderly in complex matters, diligent in seeking relevant information, reasonable in the selection of criteria, focused in inquiry, and persistent in seeking results which are as precise as the subject and the circumstances of inquiry permit. Thus, educating good critical thinkers means working toward this ideal. It combines developing CT skills with nurturing those dispositions which consistently yield useful insights and which are the basis of a rational and democratic society.*

It is worth noting that this definition has two dimensions to it: cognitive skills and affective dispositions. Facione also provides a detailed taxonomy of skills and sub-skills which helps to define the outer limits of Critical Thinking. However, some commentators regard the list as over-inclusive especially with regard to affective dispositions. Fisher and Scriven (1997) comment that the work is flawed in defining the Critical Thinker rather than Critical Thinking. Certainly, cognitive skills are more readily assessed than affective dispositions in traditionally styled examinations but perhaps, logically, if one wanted to assess the degree to which someone is a Critical Thinker, a personality test would be more appropriate?

Some issues in Critical Thinking literature regarding the construct and their implications for pedagogy and assessment

Thinking which is *not* Critical Thinking?

The corollary to disagreement about what *is* Critical Thinking, is differences of opinion concerning what *isn't*. There tend not to be clearly defined outer-edges of the construct. The Facione Delphi study gives some clues:

> *Not every useful cognitive process should be thought of as CT. Not every valuable thinking skill is [a] CT skill. CT is one among a family of closely related forms of higher-order thinking, along with, for example, problem solving, decision making and creative thinking. The complex relationships among the forms of higher-order thinking have yet to be examined satisfactorily.*

It may matter less to Critical Thinking teachers than to Critical Thinking test-writers as to what defines the outer limits of the discipline. Test-writers face criticisms of construct validity, for example, that their test is really testing the candidates' ideology, common or background knowledge, intelligence or creative thinking rather than, for example, inference, analysis or interpretation skills.

Critical Thinking pedagogy: separate or infused?

Not only is there some lack of clarity in the literature over what to include within a Critical Thinking curriculum, there is also some inconsistency concerning how the curriculum should be constructed. Is Critical Thinking:

(a) something which should be taught as a separate discipline, or

(b) something which is embedded or infused, either implicitly or explicitly, within other subject domains?

Whilst all Critical Thinking protagonists support the view that Critical Thinking should be part of students' educational experience, the conflict is whether its provision should be embedded in subject domains or stand alone as a separate academic discipline. Certainly, McPeck (1981) would, if anything, support the former view, asserting that:

> *To the extent that Critical Thinking is not about a specific subject, X, it is both conceptually and practically empty. The statement "I teach Critical Thinking",* simpliciter, *is vacuous because there is no generalised skill properly called Critical Thinking.*

[1] Briefly, the Delphi Method involves the formation of a panel of experts, who participate in a number of rounds of questions involving them sharing opinions. The experts can reconsider them in the light of comments offered by other experts. The overall agenda is to move towards a position of consensus (if not unanimity) on a particular subject.

However, this conflicts with the view of Fisher (2001):

> *Increasingly, educators have come to doubt the effectiveness of teaching 'thinking skills' in this way [implicitly] because most students simply do not pick up the thinking skills in question. The result is that many teachers have become interested in teaching these skills directly…taught in a way that expressly aims to facilitate their transfer to other subjects and other contexts.*

Is Critical Thinking an explicitly teachable skill or a natural disposition?

> *Most of us would claim that we can teach critical thinking, but not be too sure about whether we can change someone's personality.*
> (Fisher and Scriven, 1997)

Whilst some definitions promote Critical Thinking as an explicitly teachable skill, others make more of dispositions. For instance, Ennis's early view of Critical Thinking advocated a 'pure skills' approach, while his later work advocates a 'skills plus tendencies' position. One such tendency involves 'open-mindedness' (Ennis, 2002). As a synonym for openness, this is included as one of the five traits in the so-called 'Big 5' or Five Factor Theory of Personality (McCrae and Costa, 1996) and is widely accepted as a broad personality trait, which many view as fixed in amount or stable throughout adulthood.

McPeck's definition, 'the propensity and skill to engage in an activity with reflective skepticism' (1981), implies another disposition, akin to a 'spirit of inquiry', also present in the definitions advocated by Perkins, Jay and Tishman (1993) in their article aptly entitled 'Beyond abilities: a dispositional theory of thinking'. Interestingly, some critical thinking tendencies (e.g. open-mindedness, being questioning, observant) have some convergence with Guy Claxton's Positive Learning Dispositions (2006), that which a 'capacity to learn' comprises. Despite the use of the term 'disposition', his view is that developing (or teaching) dispositions is a fruitful endeavour. He deliberately clarifies his view of a disposition as 'merely an ability that you are actually disposed to make use of.'

Whether Critical Thinking is an explicitly teachable skill or a (fixed) natural disposition is a pertinent question, both for Critical Thinking teachers as well as people who devise and test Critical Thinking. Equally, what are the valid inferences end users might make from a score or mark obtained on a Critical Thinking Test? Assuming that one can infer that candidate Z has X amount of the ability at the moment of testing, the question is whether one believes this indicates a permanent or transient measure of that person as a Critical Thinker.

Conclusions

So, does Ebel's appraisal of Critical Thinking still hold true forty years on? Far from giving up, there has been considerable endeavour to define Critical Thinking. These attempts have certainly made the concept increasingly tangible and easier to measure, although there is still some way to go before a single definition is accepted by all. Furthermore, the introduction into the arena of over 20,000 students in about 1,000 educational institutions wishing to have their achievement in Critical Thinking certificated has added an additional dimension to Ebel's 'hard to measure' statement. Ebel was undoubtedly right – Critical Thinking is difficult to define satisfactorily and hard to measure. But we have not given up trying.

References

Benderson, A (1990). *Focus. Critical Thinking: Critical Issues*. Princeton: Educational Testing Service.

Claxton, G. (2006). *Expanding the Capacity to Learn: A new end for education?* Opening Keynote address at British Educational Research Association Annual Conference, September 2006.

Ebel, R.L. (1965). *Measuring Educational Achievement*. New Jersey: Prentice Hall.

Ennis, R.H. A concept of Critical Thinking. Quoted in Siegel, H. (1988). *Educating Reason: Rationality, Critical Thinking and Education*. London: Routledge.

Ennis, R.H. (1996). *Critical Thinking*. New York: Prentice Hall.

Ennis, R (2002). http://faculty.ed.uiuc.edu/rhennis/SSConcCTApr3.html accessed on 06/06/2006.

Facione, P.A. (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction: Executive Summary, The Delphi Report*. Millbrae, CA: California Academic Press.

Fisher, A. (2001). *Critical Thinking: An Introduction*. Cambridge: Cambridge University Press.

Fisher, A. & Scriven, M. (1997). *Critical Thinking: Its definition and assessment*. Norwich: Centre for Research in Critical Thinking.

McCrae, R.R. & Costa, P.T. Jr. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives*. 51–87. New York: Guilford.

McPeck, J.E. (1981). *Critical Thinking and Education*. Oxford, UK: Martin Robertson.

Paul, R. (1992). Critical Thinking: What, Why and How? *New Directions for Community Colleges*, **20**, 1, Spring 1992.

Perkins, D.N., Jay, E. & Tishman, S. (1993). Beyond Abilities: a dispositional theory of thinking. *Merril Palmer Quarterly*, **39**, 1–21.

Sternberg, R.J. (1986). *Critical Thinking: Its Nature, Measurement and Improvement*. ERIC Document Reproduction Service ED272882.

Difficulties in evaluating the predictive validity of selection tests

John F. Bell Research Division

When assessments are used for selection purposes there is a need to establish their predictive validity. Although there is literature on the predictive power of school examinations, much of it fails to appreciate the complexity of the issues involved leading Wood (1991) to comment that 'the question has proved a seductive topic for statistical dilettantes'. More recently, there has been a growth in the use of tests to assist in the admissions process of universities. There are two major reasons for this growth: the need to ensure fair access and the current inability of A-levels to distinguish between high attaining candidates (Bell, 2005a).

The most selective higher education institutions have been finding that the existing school examination system is no longer providing evidence of differences in individual merit for the highest attaining candidates. An important question that is asked of selection tests is 'do they predict future performance?' Textbooks on educational measurement usually recommend assessing this 'predictive validity' by calculating the correlation coefficient between scores on the selection test and scores on an outcome variable such as degree classification, or the score on a test at the end of the first year of the degree course.

One of the most important problems associated with evaluating the predictive validity of a selection test is that the outcome variable is only known for the selected applicants. Ideally, to evaluate predictive validity a random sample of applicants would be used. There are obvious difficulties in practice (a selective university is never likely to replace an existing selection procedure with a lottery). It is almost always going to be the case that there will be rejected candidates who will not have an outcome score.

To illustrate the effect of selection, a simulated data set of one thousand applicants was created (fuller details of this data set and the analyses described here can be found in Bell 2006, *in preparation*). It was assumed that the outcome, for example an examination mark, was related to an underlying trait and that the two selection methods are also related to the trait, that is, both tests correlate positively with the activity measure and with each other. One test will be referred to as the selection test (which is being evaluated) and the other as the original method (e.g. examination grades or interviews scores).

Table 1 : Correlations between selection methods and outcome

	Selection Test	*Original Method*	*Outcome*
Selection test	1.00		
Original method	0.28	1.00	
Outcome	0.56	0.54	1.00

The correlations in Table 1 have been set at what can be considered to be a realistic level. There are many factors that can determine outcomes in the real world that are not measured by any one test (indeed some influences can be the results of events that occur after the applicant has been admitted). The low correlation between the two selection methods indicates that they measure different traits and that both are important predictors.

There are a number of different types of selection procedure. The first type is a simple lottery, referred to as RANDOM. When lotteries have been used for selection they have either been used with other methods, either in the form of weighted lotteries, for example Dutch medical school admissions (ten Cate and Hendrix, 2001), or one stage in a medical admission (lotteries are used at one UK medical school to reduce the number of applicants to a manageable size).

The next type uses only the original method. This involves taking the *n* highest scoring applicants on the original method where *n* is the number of available places (taking the best *n* applicants is assumed for all the remaining rules). This is the situation when a selection test is being piloted so it is referred to as a PILOT because it corresponds to a pilot year where the results of the selection test play no part in admissions decisions.

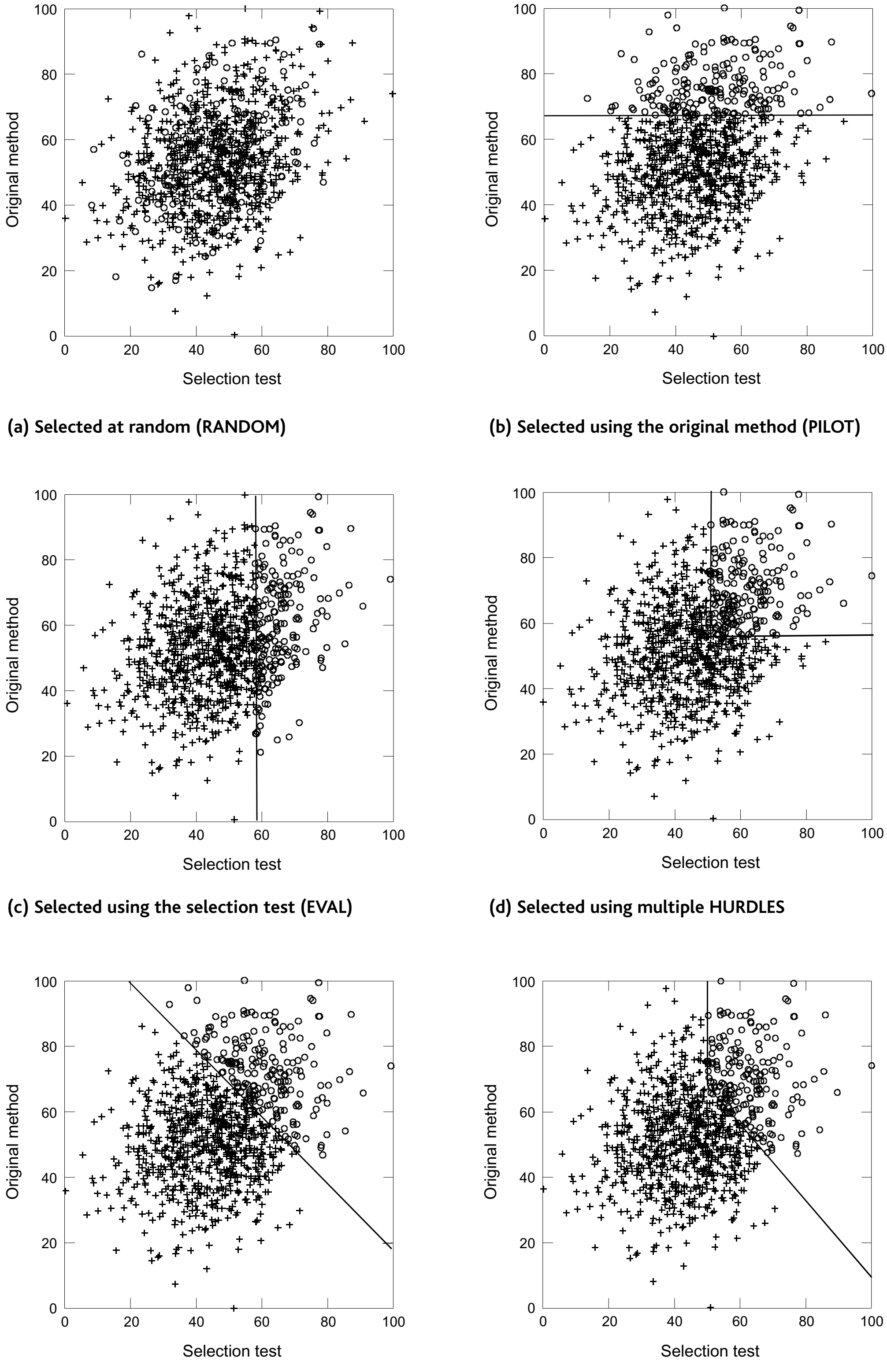
The next method will be referred to as EVAL and involves only using the selection test and ignoring the original method. This would represent the situation when a test that is the sole method of selection is being evaluated. Both PILOT and EVAL are examples of single hurdle rules.

The remaining methods involve combining test scores. The first uses multiple hurdles and will be referred to as HURDLES. This involves selecting a fixed proportion of the entry with one test (e.g. the top 40% on the selection test) and then repeating this with another test (taking the 50% with the highest scores on the original method from the top 40% on the selection test). Multiple hurdles can be used when using all the selection methods on all applicants is prohibitively expensive so the first test is used to reduce the number of applicants for the second assessment. In this case, there are obviously multiple rules that could be applied depending on the percentages used for the first hurdle.

The next method of combining scores is compensatory and will be referred to as COMPEN. This involves taking a weighted sum of the scores. In this article, equal weights have been used but obviously others could be used. The effect of changing the weights is to change the slope of the line in panel (e) of Figure 1. In a compensatory method a very poor performance on one test can be compensated by a very good performance on another. This differs from the multiple hurdles method which guarantees a level of performance on all tests.

Finally, there are hybrid methods which use both hurdles and compensation and will be referred to as HYBRID (Figure 1(f)). These are probably the most realistic in practice (e.g. a University admissions decision might depend on obtaining at least a grade B for a particularly relevant subject – a hurdle – and exceeding a particular UCAS score – a compensatory method). In the example used in this article, a hurdle is set taking the top 40% using the selection test and then the top 20% using the compensatory rule described above.

Figure 1 :
Types of selection method



(a) Selected at random (RANDOM)

(b) Selected using the original method (PILOT)

(c) Selected using the selection test (EVAL)

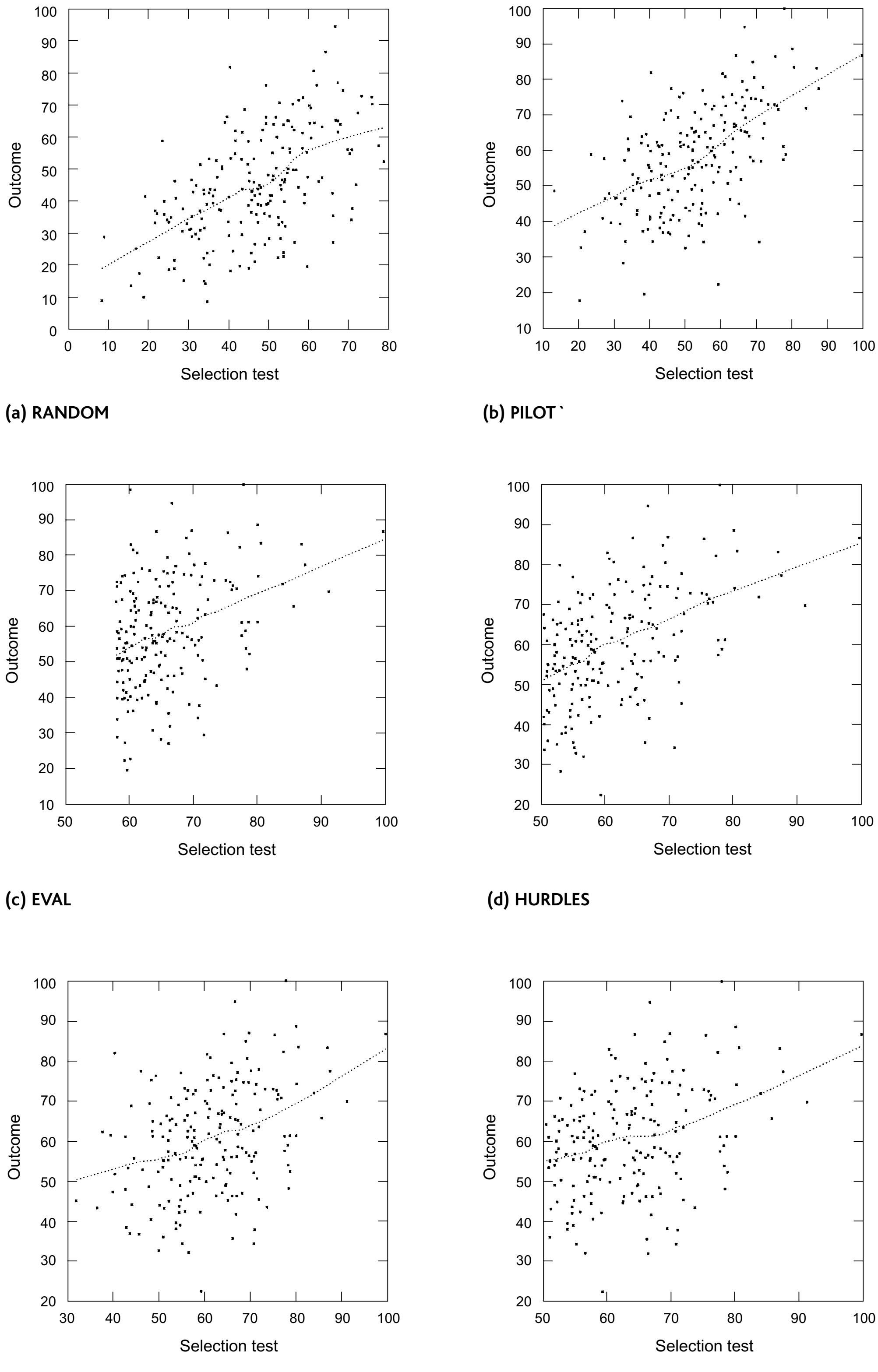
(d) Selected using multiple HURDLES

(e) Selected using COMPENsation method

(f) Selected using HYBRID method

In addition, two other rules (RANORIG and RANCOMP) were defined for comparative purposes (these are not illustrated). In this case, it is assumed that it is only possible to obtain scores on the original method for 40% of the applicants. Rather than selecting the 40% with the selection test, this selection is used at random. These rules have been defined so that the outcomes can be compared with the multiple hurdle and the hybrid rules. The first rule is a random selection followed by the original method (the graph would be like Figure 1(b) but with fewer points and a line defined by a lower pass score because there are fewer candidates to select from) and the second is a random selection followed by the compensation method (like Figure 1(e) but with fewer points and the line closer to the origin). This is sometimes proposed as a solution when there are too many applicants to interview.

Note that the last five methods are examples from families of rules defined by the choice of weights and cut scores. This means that in the following discussion conclusions about the differences between these methods should be treated with care because they might not be using the optimal version of each rule.



(a) RANDOM

(b) PILOT`

(c) EVAL

(d) HURDLES

(e) COMPEN

(f) HYBRID

Figure 2 :
Results of the selection

In the real world, only outcome data for the selected applicants is available. Scatter plots for the selected candidates are presented as Figure 2. Each part figure consists of a scatter plot of outcome against selection test for the selected applicants with a lowess smoothed line added. An inspection of the figures suggests that there is considerable variation in the strength of the relationship depending on the selection method used (this is most noticeable in the increasing spread of points even allowing for the changes in the axes).

Table 1 : Comparison of different selection methods

	Statistics			*Correlations*			*Grades*				
Method	*N*	*Mean*	*Sd*	*X1*	*X2*	*Xmean*	*A*	*B*	*C*	*D*	*E*
All	1000	45	17	0.56	0.54	0.69	20	40	60	80	100
RANDOM	200	45	17	0.55	0.64	0.72	23	37	59	80	100
PILOT	200	58	15	0.55	0.27	0.57	48	70	88	99	100
EVAL	200	58	15	0.35	0.52	0.55	45	73	87	96	100
COMPEN	200	60	14	0.49	0.31	0.51	50	79	92	99	100
HURDLES	200	60	14	0.49	0.31	0.51	50	79	92	99	100
HYBRID	199	60	13	0.35	0.28	0.47	52	80	94	99	100
RANORIG	199	53	15	0.55	0.39	0.61	34	59	80	95	100
RANCOM	200	54	14	0.32	0.33	0.48	36	62	80	86	100

(Note some rules involved ties so fewer than 200 were accepted)

In Table 1 some summary statistics about the different selection methods have been presented: the number selected, the mean and standard deviation of scores on the outcome variable for the selected applicants, correlations of the outcome with the selection test (X1), the original method (X2) and the mean of X1 and X2 (Xmean) respectively, and finally, the remaining five columns show the cumulative grade distribution for the selected applicants by dividing all the candidates into five equally sized groups based on the outcome scores. Thus, the mean score for the whole entry of 1,000 applicants is 45 with standard deviation of 17 and the correlations with three selection measures are 0.56, 0.54 and 0.69. By definition, 200 applicants in the whole entry obtained a grade A so a perfect selection method would give 100% A grade applicants. Inspecting the table reveals that the three methods that combine scores are the most successful at selecting good candidates (note it would be wrong to draw a general conclusion because neither cut-scores nor weights have been optimised). It is important to note that the predictive validity as measured by the uncorrected correlation coefficient declines as the selection methods become more effective.

Clearly, considering the correlation without considering the selection process can be very misleading. Suppose that the administration of an institution using the hybrid method squared the correlation and then concluded that the selection test only accounted for a relatively small 12% of the variation in the outcome and so abolished the selection test. If there were no change in the entry so the selection for the next year would generate results similar to the ones generated by the PILOT method, the percentage of grade A and B students would fall from 80% to 70%. This example suggests that the effectiveness of a selection procedure is better evaluated by considering the change in performance on the outcome variable rather than the correlation between scores on the outcome variable and the selection test.

One alternative to the simple correlation is to use a corrected correlation. However, the corrections vary with selection method and the availability of data. Sackett and Yang (2000) produce a very useful review of these methods. The correction not only depends on the selection method but also on the availability of the data on the original selection methods. In all cases, assumptions are made about the performance of the rejected applicants, the shapes of the relationships and the distribution of the errors.

More recently, research has been based on the fact that a selection method can be thought of as a missing data mechanism. With selection tests data are *Missing Not At Random*, abbreviated MNAR, and the missingness mechanism is termed non-ignorable. This has been applied to research into compensatory rules. In Sweden there is a complicated higher education admittance to higher education. This compensatory system involves applicants either being admitted on the basis of an admissions test or their school leaving certificate. Gustafsson and Reuterberg (1998) investigated modelling incomplete data (Muthén, Kaplan and Hollis, 1987) and found it to be a very efficient method for estimating the predictive validity of selection tests.

So far the assumption has been that if an applicant is accepted then they will take up the place at the institution. For most institutions this is not the case, since the most able applicants, although offered places, often choose to go to another institution. This is sometimes referred to as self-selection. It can have serious consequences when evaluating selection procedures. Consider two institutions P and Q. It is assumed that institution P is trying to select the best 20% and every one offered a place will take the place. Thus the second institution (Q) is only able to select from the remaining 80% of the sample. For the purposes of discussion, results for four decision rules have been generated:

SELF1: The top 20% and the next 20% are selected by the original method (as in Figure 3(a)).

SELF2: The top 20% and the next 40% are selected by the compensatory method (as in Figure 3(b)).

RANSELF1: A random selection is made from the 80% remaining after 20% is selected by the original method.

RANSELF2: A random selection is made from the 80% remaining after 20% is selected by the compensatory method.

In Figure 3 the crosses represent the applicants selected by institution P, the circles represent the applicants selected by institution Q and the pluses represent the rejected applicants. The two selections are very different. In the first the applicants attending Q have very varied scores on the selection test but do not vary much on the score from the original method. In the second the two scores are inversely related with an applicant with a high score on the original method having a low score on the selection test and vice versa. The effect on the outcomes is that predicted gains from having a high score on the original method will be cancelled out by the losses associated with a low score on the selection test. Obviously in the real world the effect of self-selection is not so clear cut because applicants apply to more institutions and do not necessarily apply to the best institution where they could have gained a place or have taken up the place if they applied and were successful.

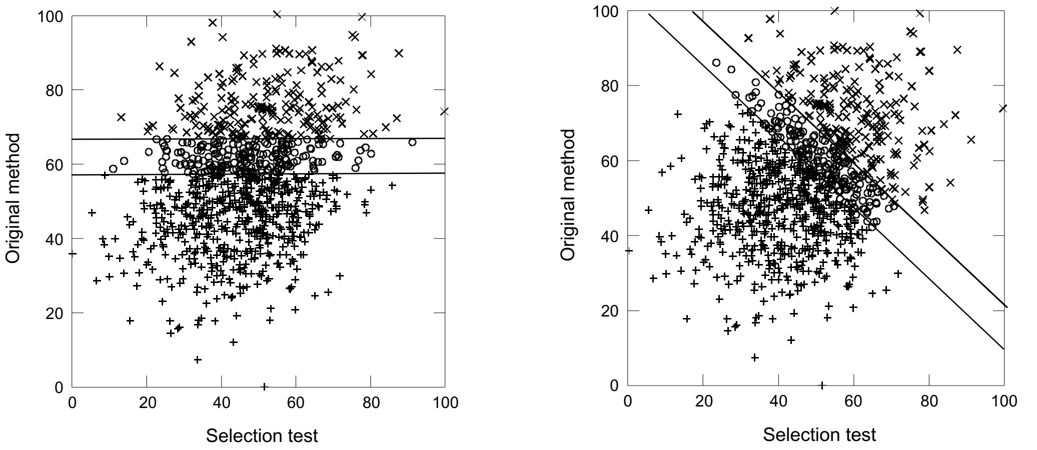


Figure 3 : The effect of self-selection on applicants accepted by two institutions

(a) Using original method for both selections (SELF1)

(b) Using compensation method for both selections (SELF2)

The last two rules serve as baselines for the first two rules. RANS1 are the results for a random sampling after institution P had selected 20% by the original method and RANS2 is the same apart from the use of the compensation rule (i.e. taking a random selection of candidates from below the upper lines in Figure 3 ignoring the lower line). The effect of the self-selection is to reduce the relative proportion of good applicants that can be selected.

Using the summary statistics in Table 3, it is clear that if the effects of selection are ignored then this could lead to serious misinterpretation of the data. For the situation described by SELF1 then it might be concluded that the new test was greatly superior to the original method. Although the correlations for the two tests are similar in the whole population, the correlation of the original methods is much greater for the selection test. For SELF2 it is possible to erroneously conclude that the selection method was ineffective and they would do better switching to a lottery. Both the correlations for candidates attending in institution Q are close to zero. However, the institution would get a much poorer entry if they did so (i.e. 27% grade A candidates for SELF2 and just 11% for RANS2). Although this example is a simulation, it is not the case that it has just been cunningly contrived to illustrate an unlikely theoretical situation; such problems occur in real life. Linn and Dunbar (1982) found that the correlations between SAT scores and subsequent performance were low for a New York community college. This was the result of students who scored highly on the SAT almost always choosing to go to better colleges.

This simulated example is obviously a gross simplification given that institutions would not necessarily use the same selection procedures and more than one institution may be involved. However, recently there has been research into applying a range restriction to situations involving institutional and self-selection. Yang, Sackett and Nho (2004) proposed a procedure using non-ignorable double selection models and found that in simulations their model produced an unbiased estimate for the population correlation.

Evaluating rules in this situation is also more complicated. If institution Q improves its selection procedure then this would not guarantee an improved entry. This is because the quality of the available applicants can also decline if institution P also improves its selection method. Such a situation would occur if both institutions introduced a selection test at the same time.

The simulated data used in this paper has demonstrated that interpreting uncorrected correlation coefficients is difficult and, depending on the circumstances, can seriously underestimate the effectiveness of a selection test. The correlation coefficient can be corrected for the effects of selection but it is important to recognise that the correction method should match the selection procedure. Unfortunately, the corrections depend on assumptions about the rejected applicants. Although it is usually argued that the assumptions made about the rejected applicants are untestable, this is not quite true. Selection tests are usually used repeatedly, meaning that the effects of changes can be monitored. For example, consider the simplest case, where in the first year the selection test is piloted but not actually used to select students. Then the range correction formula for this situation can be applied. If in the second year the entry is identical in characteristics to the first year and the selection test alone is used then a prediction of the expected correlation can be made by inverting the appropriate correction formula. This can be compared to the observed correlation.

More fundamental, however, is the question whether using the correlation coefficient in the first place as a measure of the predictive

Table 3 : The results for the self-selection rules

	Statistics			*Correlations*		*Grades*				
Method	*N*	*Mean*	*Sd*	*X1*	*X2*	*A*	*B*	*C*	*D*	*E*
All	1000	45	17	0.56	0.54	20	40	60	80	100
SELF1	202	49	14	0.51	0.20	21	54	65	81	100
SELF2	200	52	12	0.05	0.01	27	69	83	96	100
RANS1	199	43	15	0.50	0.46	16	32	47	79	100
RANS2	200	42	14	0.50	0.28	11	32	55	77	100

validity is the best basis for evaluating a selection test. The objective of the selection test is to select the students who will perform best on the outcome measures. This leads to the conclusion that it might be better to evaluate the predictive validity of a selection procedure in terms of the improvement in the quality of those selected. This could be based on a change in mean score or proportion of satisfactory students. The case of a binary outcome is discussed in more detail in Bell (2005b, c).

This article shows that it is possible that by using simplistic analyses the benefits of using selection tests may have been underestimated. For example, in the late 1960s there was an experiment using a SAT-style test in the United Kingdom (Choppin *et al.*, 1972; Choppin and Orr, 1976). The results of the experiments were considered to be something of a disappointment despite the fact that the test had been carefully designed. There was a considerable degree of selection, for example, only 26% of those who sat the test were admitted to universities. The authors of the reports used simple correlations and regression to analyse the data. It is interesting to note the patterns of results for individual institutions for mathematics. The institution with the highest mathematics scores (presumably an institution not affected by self-selection) and so a very high degree of selection, had a correlation of 0.36 for both the mathematics and verbal scores. However, the correlations were much lower and in some cases slightly negative for an institution which would have been selective and been affected by self-selection. From the simulation it is clear these results are consistent with an effective selection test, although it is also true this need not be the case. The problem is that the analyses are based on simple correlations. This is not a criticism of the authors of both reports. Both theory and the technology have advanced a long way from the 1970s. However, it is reasonable to conclude that there is a possibility that the conclusions about the ineffectiveness of this test were erroneous.

In conclusion, when a researcher makes a sweeping claim about the ineffectiveness of an admissions test but bases their argument on an uncorrected correlation or a simple regression analysis and does not consider the effects of selection, then there is a distinct possibility that such a claim is mistaken. Higher education admissions are important and it is vital that care is taken with them. Thus it is vital that research into admissions tests address in full the complexities of the data that arise from their use.

Acknowledgements

The translations from Dutch were made by a colleague, Mark Claessen.

References

Bell, J.F. (2005a). Gold standards and silver bullets: Assessing high attainment. *Research Matters: A Cambridge Assessment Publication*, **1**, 16–19.

Bell, J.F. (2005b). Evaluating the predictive validity of a selection test. Part 1 – Replacing an existing procedure. *Submitted for publication*.

Bell, J.F. (2005c). Evaluating the predictive validity of a selection test. Part 2 – Supplementing an existing procedure. *Submitted for publication*.

Bell, J.F. (2006). The effect of the selection method on the evaluation of the predictive validity of a selection test. *In preparation*.

Choppin, B.H.L., Orr, L., Kurle, S.D.M., Fara, P. & James, G. (1973). *Prediction of academic success*. Slough: NFER Publishing.

Choppin, B. & Orr, L. (1976). *Aptitude testing at eighteen-plus*. Slough: NFER Publishing.

Gustafsson, J.-E. & Reuterberg, S.-E. (2000). Metodproblem vid studier av Högskole-provets prognosförmåga – och deras lösning. [Methodological problems in studies of the prognostic validity of the Swedish Scholastic Aptitude Test (SweSAT) – and their solution] *Pedagogisk Forskning i Sverige*, **5**, 4, 273–284. (In Swedish with extensive English summary)

Linn, R.L. & Dunbar, S.B. (1982). Predictive validity of admissions measures: correction for selection on several variables. *Journal of College Student Personnel*, **23**, 222–226.

Muthén, B., Kaplan, D. & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, **42**, 431–462.

Sackett, P.R. & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, **85**, 112–118.

ten Cate, T.J.T. & Hendrix, H.L. (2001). De eerste ervaringen me slectie [Initial experience with selection procedures for admission to medical school]. *Nederlands tijdschrift voor geneeskunde*, 14 Juli:145, 28, 1364–1368.

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

Yang, H., Sackett, P.R. & Nho, Y. (2004). Developing a procedure to correct for range restriction that involves both institutional selection and applicants' rejection of job offers. *Organisational Research Methods*, **7**, 4, 442–455.

PREDICTIVE VALIDITY

Using Thinking Skills Assessment in University admissions

Joanne Emery and John F. Bell Research Division

In the first issue of *Research Matters*, the difficulties involved in assessing high attaining candidates were discussed (Bell, 2005a). A particular problem is that elite institutions are faced with selecting among candidates with the same grades on existing qualifications. Most applicants to the University of Cambridge are predicted, or have already, at least three grade As at A-Level. Cambridge University admissions staff therefore requested that Cambridge Assessment (then known as UCLES) develop a 'Thinking Skills Assessment' (TSA) to assist in making admissions' decisions. When first proposed, the TSA was seen as a test that would form part of the admissions interview process so that it could be taken by applicants during their interview visits to Cambridge. This has the advantage in the Cambridge context of allowing the use of the test on a college-by-college and a subject-by-subject basis. At the time of writing, most Cambridge colleges use the TSA during the admissions process and the range of subjects for which it is used varies from college to college. The test provides *supplementary information* for use in helping to make admissions decisions. Obviously, to be meaningful, any such selection tool must be able to predict future performance. This issue of predictive validity is the focus of this article.

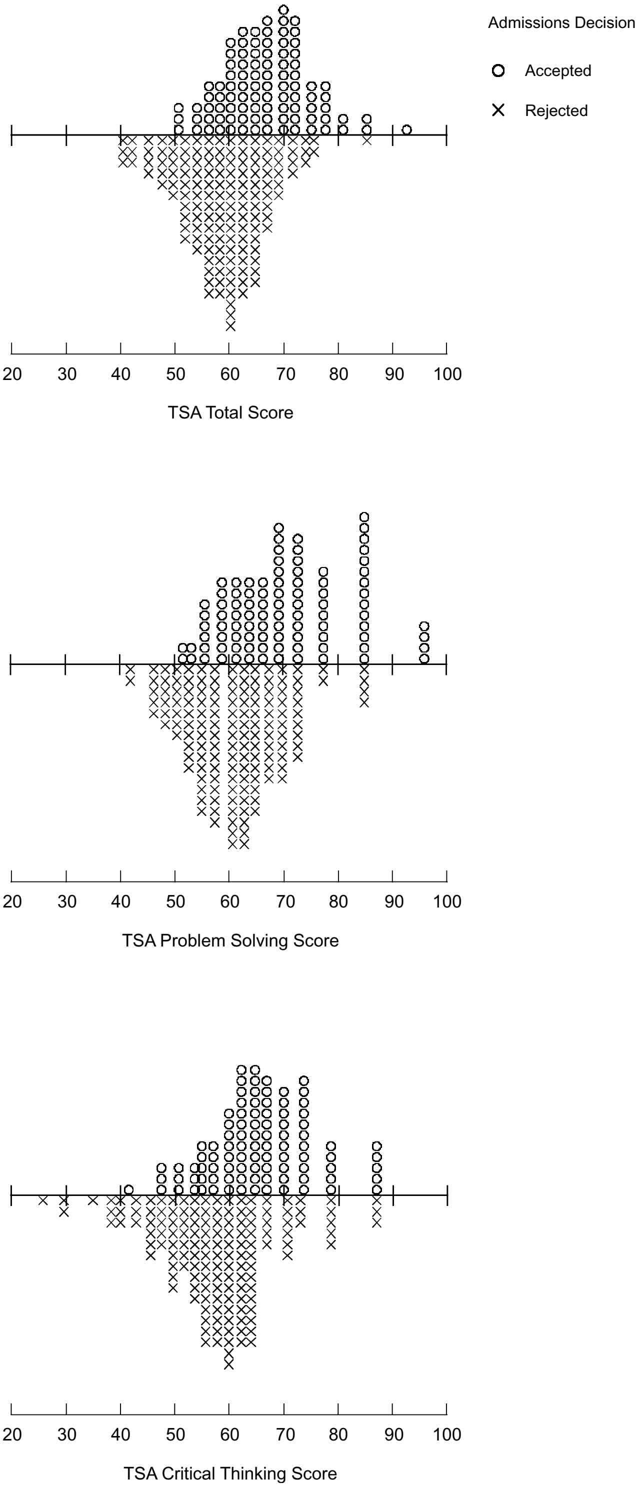
The Cambridge Thinking Skills Assessment (TSA) provides an assessment of two kinds of thinking: Problem Solving and Critical Thinking. Problem Solving describes reasoning using numerical, graphical and spatial skills. It requires developing strategies for tasks through thought and planning. Critical Thinking is often defined as 'reasonable, reflective thinking that is focussed on deciding what to believe or do' (Ennis, 1996). Central to Critical Thinking are the concepts of argument and evaluation. It requires the ability to interpret, summarise, analyse and evaluate arguments and ideas. With the TSA, the aim is to provide an assessment of *Thinking Skills*: intellectual skills that are independent of subject content and are generalisable to a wide range of subject areas. For example, the skill of Critical Thinking can be useful in subject areas ranging from the Humanities (interpreting documents and evaluating their arguments) and the Arts (following the reasoning of great thinkers) right through to the Sciences (appreciating advances in scientific development).

Cambridge Assessment has a long history of developing tests under the general heading of 'Thinking Skills'. An item bank of former Thinking Skills questions (items) was built up for this purpose. This gave an excellent starting point for the development of the TSA. The test consists of 50 multiple-choice questions, each with 5 possible answers, and has a time limit of 90 minutes. Questions assessing Problem Solving and Critical Thinking skills are mixed throughout the test and there are no penalties for incorrect responses. In December 2001, 289 Computer Science applicants took the TSA. This expanded to 472 in December 2002 with more colleges and more subjects taking part.

Up to this point the objective of the TSA work was the development and evaluation of the test itself but in January 2003 Cambridge Assessment added a second objective: that of experimental online delivery of the test. This software was developed specifically for Cambridge Assessment as a prototyping system. Both objectives were successfully achieved: there was a greatly enhanced take-up of the test, with 23 colleges taking part involving 4 main subjects (Computer Science, Engineering, Natural Sciences and Economics), and the administration procedures were based around the online system we had developed. A total of 1,551 tests were administered in that year: 1,114 paper tests and 437 online tests. An especially valuable feature was the administrative website used for making entries (registration) and returning results. Online tests were marked automatically and paper tests were marked using scanning technology with intelligent character recognition. A website (http:// tsa.ucles.org.uk) is available giving details of the TSA with example and practice materials.

This article reports on the 2003 TSA scores and the subsequent 1st year (Part 1A) examination results of Computer Science students (taken in Summer 2005). Of the 1551 candidates who sat the TSA in 2003, 238 applied to study Computer Science. Of these, 94 received an offer of a place and 144 were rejected. TSA scores are reported as a total calibrated score and as separate calibrated Problem Solving and Critical Thinking sub-scores. The calibration process allows the results of candidates taking different versions of the test to be reported on a common interval scale.

Figure 1 : Dot density plots showing the TSA 2003 score distributions of candidates who were offered a place (conditional or unconditional) and candidates who were rejected for the Computer Science course



Plots showing the TSA score distributions of Computer Science candidates who were offered a place (conditional or unconditional) and candidates who were rejected are shown in Figure 1. These plots are particularly helpful in evaluating whether the TSA is likely to be useful as a preliminary hurdle to reduce the number of interviews given (at the time of writing it is rare for an applicant not to be interviewed). It can be seen that few of the accepted candidates had low scores. If the test were to be used for pre-selection, two questions need to be considered: why any relatively low-scoring candidates were accepted and whether they could be spotted without an interview.

Correlations, whilst problematic, are the most familiar measure of predictive validity. Table 1 displays the correlation coefficients between the 2003 TSA scores and 1st year Computer Science examination performance in 2005. Pearson coefficients are given for all variables except 'rank' where a Spearman's rho is used due to the ordinal nature of the data. It should be noted that the coefficients displayed are uncorrected for the effects of selection. Correlations tend to produce underestimates where selection tests are concerned due to restricted score ranges. Although there are corrective formulae, none of them apply to this particular situation where the selection test is used in conjunction with other qualitative information. There are, however, some guidelines that can be applied.

General guidelines for interpreting validity coefficients

Validity coefficient value	*Interpretation*
above .35	very beneficial
.21–.35	likely to be useful
.11–.20	depends on circumstances
below .11	unlikely to be useful

(US Department of Labor, Employment Training and Administration, 1999)

Table 1 : Correlations between TSA 2003 scores and Part 1A examination outcome in Computer Science

	N	*TSA Total Score*	*TSA Problem Solving Score*	*TSA Critical Thinking Score*
Computer Science Part 1A Rank in Year	64	–.453**	–.439**	–.292**
Computer Science Part 1A Total Mark	64	.445**	.419**	.315**
Computer Science Paper 1 Mark	67	.488**	.477**	.327**
Computer Science Paper 2 Mark	64	.566**	.505**	.425**

** The correlation is significant at the 0.01 level (one-tailed)

The total TSA score and both the Problem Solving and the Critical Thinking components show highly significant positive correlations with 1st year examination performance in Computer Science. The relationships are slightly stronger for the Problem Solving component than the Critical Thinking component but show the greatest magnitude for the combined total score. Paper 1 of the examination covers topics on the Foundations of Computer Science, Operating Systems, Algorithms and Java Programming. Paper 2 is composed of questions on Digital Electronics and Discrete Mathematics.

Table 2 displays the means and standard deviations of the TSA scores of candidates achieving various Part 1A examination classes in 2005. Total examination marks are graded (in descending order of merit) as class 1, class 2:1, class 2:2, class 3, ordinary and fail. Students obtaining 1st class results tend, on average, to have gained higher total scores on the TSA than those who went on to obtain a 2:1, who, in turn, tend to have achieved higher scores than those obtaining a 2:2. This is also the case for the Problem Solving and Critical Thinking sub-scores. It is notable that the average TSA scores of students gaining a 3rd class outcome are higher (for total score and Critical Thinking) than those of candidates gaining a 2:2. It is quite likely that candidates who obtain such poor results do so for reasons that are not necessarily related to their academic abilities, given that they have very high academic achievement prior to their arrival at Cambridge. An admissions test can only identify those students who are capable of doing well: not necessarily those who will do well.

Table 2 : TSA descriptive statistics by examination class achieved in Computer Science

Part 1A Class	*N*	*TSA Total Score*	*TSA Problem Solving Score*	*TSA Critical Thinking Score*
Class 1	16	71.5 (7.99)	75.2 (12.09)	70.2 (8.70)
Class 2:1	18	68.5 (6.96)	74.6 (11.46)	65.8 (9.43)
Class 2:2	23	63.4 (7.36)	66.2 (8.60)	61.4 (8.32)
Class 3	7	63.9 (6.14)	62.7 (7.18)	65.7 (6.63)

The weakness of correlation analysis here is that it cannot include data for candidates who have been rejected. The TSA is used in a complex process which is compensatory in nature but not necessarily quantified. This means that there is no simple way of adjusting the coefficients for selection effects. However, there is an alternative method of evaluating predictive validity. When a selection procedure is based on the principle of maximising academic performance then this is the same as assuming that, for a given TSA score, the probability of obtaining a degree of a particular class is lower for applicants who were rejected compared with those who are accepted. There is no way of directly testing this. However, tau analysis has been developed to investigate this assumption (Bell, 2005b, 2005c).

The tau method uses logistic regression equations to calculate the probability that any given TSA score will result in the student who achieved it gaining a 1st class result. The students who were actually selected for course entry using the existing methods (predicted grades, interview performance, UCAS form information) are compared to the students who *would* have been selected if TSA scores alone had been used, in terms of how many 1st class outcomes they achieved (or *would* have achieved). The probable number of 1sts which would have been achieved with the TSA-only method is calculated by taking the top *n* highest-scoring TSA candidates (the same number as were actually selected) and simply summing together their calculated probabilities of success.

The above analysis requires assumptions to be made about the probabilities of success for the rejected applicants. The magnitude of these probabilities is related to the degree of confidence in the existing selection system. It is assumed that, for any given mark on the TSA, any candidate who was rejected by the existing system had a lower probability of success than one who has been accepted. The tau method quantifies this difference. This is achieved by multiplying the probabilities for the rejected candidates by a confidence factor (1 minus 'k'). The possible values of k can range from 0 (no confidence) to 1 (absolute confidence). Absolute confidence implies that the procedure has definitely selected the best candidates and no confidence suggests the selection procedure was essentially random (this is not plausible if the logistic regression for selected applicants is positive). In practice, the confidence level is not known. However, it is possible to investigate the predictive validity of the test by considering a range of confidence values.

Table 3 : Tau analyses comparing the probable success rates achieved using new (TSA-only) versus existing selection methods

COMPUTER SCIENCE PART 1A 2005 — *confidence in existing system: k=0.75*

	New Predictor Measure		
	TSA Total Score	*TSA Problem Solving Score*	*TSA Critical Thinking Score*
Number of applicants	210	210	210
Number accepted	67	67	67
Number selected with new method	67	70	67
Actual number of firsts achieved	16	16	16
Predicted firsts for new method	16.2	16.3	16.4
Observed success rate	0.24	0.24	0.24
Predicted success rate of new method	0.24	0.23	0.24
new/existing	1.0	1.0	1.0
new/random	2.5	2.3	2.4
existing/random	2.4	2.3	2.3

confidence in existing system: k=0.5

	New Predictor Measure		
	TSA Total Score	*TSA Problem Solving Score*	*TSA Critical Thinking Score*
Number of applicants	210	210	210
Number accepted	67	67	67
Number selected with new method	68	67	67
Actual number of firsts achieved	16	16	16
Predicted firsts for new method	17.2	16.4	17.7
Observed success rate	0.24	0.24	0.24
Predicted success rate of new method	0.25	0.24	0.26
new/existing	1.1	1.0	1.1
new/random	2.1	1.9	2.1
existing/random	2.0	1.8	1.9

(Note: any differences between the numbers accepted and numbers selected with the new method are due to tied ranks in TSA scores)

The tau tables presented above show the case where k=0.75 (thus assuming high confidence in the existing system) and the case where k=0.5. The 'accepted' candidates are those who were actually selected by the colleges and for whom the number of 1st class results is known. The 'selected' group are those who would have been chosen on the basis of the TSA alone (the total score or its subscales). The 'random' group is akin to using a lottery method: its success rate considers the probable number of 1st class results for the entire applicant pool. The success rates and ratios presented above compare the proportion or likely proportion of students gaining a 1st class outcome using the old, new and random selection methods.

The results suggest that, even when confidence in the existing system is high, using the total TSA score alone would have resulted in at least the same success rate as was obtained using the existing selection methods. If confidence is lower (k=0.5) then the total TSA score and the Critical Thinking component both outperform the existing selection system. However, it is not necessarily the case that the same candidates would have been accepted. The comparison of existing versus new methods shows the effect of changing from using all the information, including the TSA, to using the TSA alone. The success rates for both methods, however, are vastly superior to a random selection of candidates from the applicant pool.

In this article we have demonstrated that a Thinking Skills Assessment is useful in the University admissions process as an additional source of evidence. Correlations with subsequent examination performance are impressive, given the problems of restricted score ranges in such highly selected candidates. Students attaining higher examination classes tended to have achieved higher TSA scores and the tau analyses suggest that selecting on the basis of the TSA alone would have produced at least the same number of Class 1 results. In conclusion, there are substantive differences in Thinking Skills between candidates with three grade As at A-Level and these differences predict their future performance. Thus a selection process involving the assessment of Thinking Skills is necessary.

References

Bell, J. F. (2005a). Gold standards and silver bullets: Assessing high attainment. *Research Matters: A Cambridge Assessment Publication*, **1**, 16–19.

Bell, J. F. (2005b). Evaluating the predictive validity of a selection test. Part 1 – Replacing an existing procedure. *Submitted for publication*.

Bell, J. F. (2005c). Evaluating the predictive validity of a selection test. Part 2 – Supplementing an existing procedure. *Submitted for publication*.

Ennis, R.H. (1996). *Critical Thinking*. New York: Prentice-Hall.

US Department of Labor, Employment Training and Administration (1999). Employer's guide to good practices. (accessed at: http://www.chrysaliscorporation.com/pdf/Testing_and_Assessment_Guide.pdf)

Factors affecting examination success at A-level

Carmen L. Vidal Rodeiro and John F. Bell Research Division

Introduction

Previous research has shown that background information about students (such as gender or ethnicity) is an important predictor of attainment (e.g. Gray *et al*. 1990, Haque and Bell 2001, Bell 2003, OECD 2004 or Raffe *et al*. 2006). This previous research has also provided evidence of links between socio-economic characteristics of students and their educational attainment, for example, measures of socio-economic status, parents' educational background, family structure and income have been shown to be important predictors of attainment at secondary level. Such factors have also been found to be strongly related to measures of prior attainment at entry to school.

In this research we are going to use information from different databases in order to investigate the contribution of students' attainment at GCSE, family background, schooling and neighbourhood to their success in GCE A-levels. We will focus on the students' performance in GCE A-level in Chemistry.

Data

Data on students' examination results for the cohort of students that were 17 years old in 2004 were used. These data have been combined with the National Pupil Database (NPD) which incorporates ethnic group, first language, free school meals eligibility (FSM) and special education needs (SEN). A description of the NPD data is given in Vidal Rodeiro (2006).

The inclusion of students' previous attainment as an explanatory variable in a model allows the investigation of the effect of background factors on relative levels of attainment. The prior attainment of the students was based on the mean of their GCSE results using the usual points scale (A*=8, A=7, B=6, etc).

School characteristics were derived using data from the awarding bodies' national centre database and the 16+/18+ databases for the entire A-level entry in England in 2004. Schools offering GCE A-level subjects were classified into five categories: comprehensive and secondary modern schools, further education (FE) and tertiary colleges, grammar schools, independent schools and sixth form colleges. In addition, the attainment group of the schools was computed as the mean of the attainment of their students which was based on their A-level points score. To compute this score, all students in the 2004 cohort with at least three A-level results, excluding general studies, were selected. The A-level grades for these students were converted into points using the UCAS old tariff (A=10, B=8, C=6, etc) and the sum of the points of their three best A-levels was computed. Schools were then allocated into five attainment groups.

Table 1 shows the attainment group by school type. Around 83% of the grammar schools and 89% of the independent schools are in the highest attainment group. This compares to the 10% of comprehensive schools or the 17% of sixth form colleges. No FE/Tertiary colleges are in the highest attainment group.

Table 1 : School attainment group by school type (column percentages)

School Attainment Group	*Comprehensive schools*	*Grammar schools*	*Independent schools*	*Sixth form colleges*	*FE/Tertiary colleges*
Group I (Low)	4.2	0.0	0.2	0.4	17.2
Group II	14.4	0.1	0.8	10.5	46.9
Group III	29.5	1.3	2.8	44.3	21.2
Group IV	41.8	15.2	7.7	28.1	14.6
Group V (High)	10.1	83.4	88.6	16.7	0.0

A female ratio per school was computed (number of females taking A-levels in the school over the total number of students in the school; see Malacova, 2006, for details). If the female ratio was 1, the school was considered a 'Girls only' school. If the female ratio was 0, then the school was considered a 'Boys only' school. The rest of the schools were considered coeducational or mixed schools. Sixth forms were also classified into five groups according to their size (based on the number of students in the upper sixth form): less than 30, 30 to 59, 60 to 119, 120 to 239, and more than 240.

Not everything that might have an influence on the students' success in a particular examination is the result of their previous attainment and the school characteristics. Students' motivation and subject preference, for example, might be important too. Further research is being carried out by Cambridge Assessment about subject choice and motivation.

Recent studies have found that neighbourhood-level variables have an important influence on educational attainment (e.g. Ensminger *et al*., 1996, OECD, 2004, Raffe *et al*., 2006). In this research, the characteristics of the neighbourhood in which a school is situated are considered. There is a risk that the address of a school may not reflect its catchment area. For example, a school might be located near the boundaries of a ward thus attracting a large proportion of children from other wards, or a school could have been affected by parental choice. These problems could be removed if it were possible to use the postcodes of the students' home address (instead of the postcodes of the schools), in conjunction with the ward level census data, but these data were not available to us. Despite these limitations, significant correlations can be identified between school examination performance and various indicators derived from the ward level census data.

Data about electoral wards in England were obtained from the Neighbourhood Statistics Service managed by the Office of National Statistics and it was matched to the postcodes of the schools. In this research we focus on the following factors: parental unemployment, parental qualifications, car ownership, density of population (proxy for rural/urban areas), lone parent status, ethnicity and deprivation index.

Methods

A multilevel modelling technique was used. Multilevel models allow for the clustering of individuals within schools and they do not violate the assumption of independence of observations that traditional ordinary least squares analysis commits when analysing hierarchical data. For example, individual students are grouped into schools; students in the same school may have more in common than with students in other schools. Multilevel models take account of this hierarchical structure of the data and produce more accurate predictions.

The modelling process was conceived as a two-level model in which students (level 1) were nested in schools (level 2). The explanatory variables (prior attainment, gender, school characteristics, etc.) were entered into the fixed part of the model. The outcome measure is the attainment at the completion of the A-level stage.

The models were fitted using the programme MLwiN (Rasbash *et al*., 2005). They were run for various combinations of students and school characteristics and background and socio-economic factors.

Results

The total number of students obtaining an A-level in Chemistry in 2004 was 27,867. More than 50% of these students obtained at least grade B. 32% of these students obtained grade A. Only 5% failed to obtain at least grade E. Table 2 shows the number and percentage of students per school type and school gender that obtained an A-level in Chemistry and Table 3 shows the grade distribution by type of school.

Table 2 : Number of A-level Chemistry students by school type and school gender

		Number of students	*Percentage of students*
School Type	Comprehensive	10722	38.5
	Grammar	3773	13.5
	Independent	6349	22.3
	Sixth Form	1635	5.9
	FE/Tertiary	885	3.2
School Gender	Boys	2180	7.8
	Girls	3478	12.5
	Coeducational	17921	64.3

Percentages shown in Table 3 are column percentages. Among all the students that obtained grade A in their Chemistry A-level, around 42% studied in an independent school, 20% in a grammar school, 6% in a sixth form college and 31% in a comprehensive school. From this table, it is possible to see that there are differences in the performance of students by type of school but do they disappear when we adjust for other factors, such as students' attainment?

We first studied students' characteristics. Secondly, additional models were fitted for various combinations of school characteristics. Finally, characteristics of the neighbourhood where the schools were located were introduced.

In the following we will report the results obtained when only students obtaining grade A and at least grade E were considered.

Table 3 : A-level Chemistry grade distribution by type of school (% from each school type obtaining each grade)

School Type	*Grade*					
	A	*B*	*C*	*D*	*E*	*U*
Comprehensive	31.4	44.9	53.7	62.5	66.0	72.4
FE/Tertiary College	0.8	1.6	2.3	2.9	3.2	4.4
Grammar	19.6	18.3	15.2	12.3	11.4	8.8
Independent	42.5	27.8	20.5	14.6	11.7	6.6
Sixth Form College	5.7	7.4	8.3	7.7	7.7	7.8

Grade A

For these analyses, the dependent variable takes the value 1 if the student obtained grade A and 0 otherwise.

The results of a first model which included the gender of the student, the prior attainment (in all the analyses the mean GCSE was centred on its mean value of 6.77) and their interaction are reported in Table 4. If *b* is the logistic regression coefficient for a particular variable (estimate), then exp(*b*) is the odds ratio. The odds ratio for each independent variable gives the relative amount by which the odds of obtaining a grade A increase (O.R. greater than 1) or decrease (O.R. less then 1) when the value of the independent variable is increased by one unit.

For example, the variable 'male' is coded as 0 (=female) and 1 (=male) and the odds ratio for this variable is 2.2. This means that the odds of males obtaining grade A are 2.2 times higher than the odds of females.

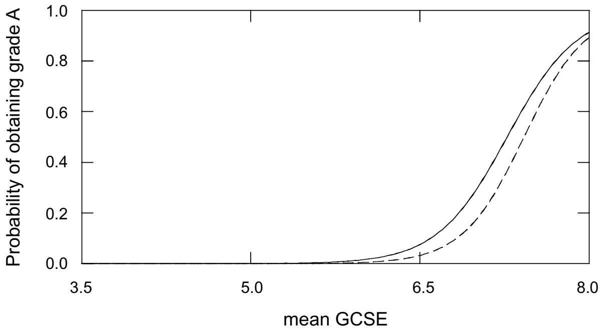
Table 4 : Individual characteristics I[1]

	Estimate	*Standard Error*	*Odds Ratio*
Constant	-1.914	0.044	
Male	**0.806**	0.053	**2.2**
Mean GCSE	**3.234**	0.066	**25.4**
Male*Mean GCSE	**-0.435**	0.087	**0.6**
School-level variance	**0.406**	0.037	

[1]. Estimates in bold indicate statistical significance at 0.05 level.

Figure 1 shows the predicted probability of obtaining grade A by mean GCSE generated by the estimates in Table 4. Although, on average, male students are less likely to obtain a grade A for any given value of mean GCSE, the difference is smaller for the most able male students. There are two difficulties with interpreting these data. First, relative progress is being considered and the sex difference can be related to the mean GCSE performance or to the A-level performance or to both. Second, there are

Figure 1 : Predicted probability of obtaining grade A by mean GCSE (solid line for females and dashed line for males)



selection effects, for example, it is possible that the motivation of students differs between groups.

There was a large number of students who had their ethnic group and other personal data missing (e.g. PLASC data does not include results from some independent schools or non-maintained special schools). In order to study the effect of the ethnicity we used a reduced data set (15,613 students, 56% of the original data), where data about ethnicity, first language, free school meals eligibility and special education needs were available.

Substantial differences appeared between ethnic groups (Table 5). Of course, these differences could be the result of other variables that have not been included in the model but vary by ethnic group. Additionally, some of the ethnic groups are very broad and if they were split the results could differ. After controlling for students' attainment, the results show that in comparison to the 'white' group, Bangladeshi, African, Chinese or Indian students have a higher probability of obtaining grade A.

Table 5 : Individual characteristics II

	Estimate	*Standard Error*	*Odds Ratio*
Bangladeshi	**0.438**	0.142	**1.5**
African	**0.570**	0.194	**1.8**
Caribbean	-0.230	0.517	0.8
Chinese	**0.447**	0.185	**1.6**
Indian	**0.678**	0.110	**2.0**
Mixed	**0.609**	0.162	**1.8**
Other ethnic group	**0.435**	0.132	**1.5**
Language – not English	-0.084	0.100	0.9
FSM	-0.292	0.170	0.7
SEN	0.178	0.444	1.2

Next, school type, school gender, school attainment and school size were included in the model as sets of dummy variables. Comprehensive, coeducational, attainment group 1 and size 1 schools were assigned the baseline. Since prior attainment was significant as an individual factor of examination success, it was also included. Results for this model are displayed in Table 6.

The odds of obtaining grade A for a student attending a grammar school are 0.9 times the odds of a student attending a comprehensive school (although this effect is not significant). However, if the student attends a sixth form college or a FE/Tertiary college, the odds of obtaining grade A are 1.8 and 1.5, respectively. Therefore, students attending sixth form or FE/Tertiary colleges have a positive advantage in their A-level Chemistry outcome.

Separate models were fitted to find out the effects of the school type when the school attainment group is not considered. In that case, the effects of grammar and independent schools on attainment in A-level Chemistry are positive and significant. There is, however, no evidence that independent schools do better on average than other types of schools once prior attainment has been taken into account.

After controlling for students' prior attainment, the school attainment group plays an important role in the success of a student taking Chemistry A-level. The higher the attainment of the school, the larger the odds of getting grade A. This supports the results found by Rutter *et al*. (1979) who reported that when students of similar prior attainment at the point of entry attending schools with differing proportions of more able students, those attending the schools with the higher percentages of more able students did better in their examinations.

Table 6 : School characteristics

	Estimate	*Standard Error*	*Odds Ratio*
Constant	-2.848	0.341	
Grammar	-0.118	0.094	0.9
Sixth Form	**0.586**	0.120	**1.8**
Independent	0.111	0.097	1.1
FE/Tertiary College	**0.414**	0.149	**1.5**
Boys school	0.134	0.091	1.1
Girls school	**-0.400**	0.077	**0.7**
Attainment 2	**0.837**	0.269	**2.3**
Attainment 3	**0.892**	0.261	**2.4**
Attainment 4	**1.109**	0.259	**3.0**
Attainment 5	**1.475**	0.265	**4.4**
Size 2	0.259	0.238	1.3
Size 3	0.221	0.224	1.3
Size 4	0.213	0.228	1.2
Size 5	0.112	0.238	1.1
Mean GCSE	**2.842**	0.044	**17.2**
School-level variance	**0.348**	0.035	

Attending a single sex school has different effects on success. The odds of obtaining grade A for a student attending a 'Boys' school are 1.1 the odds of a student attending a coeducational centre. However, the odds of obtaining grade A for a student attending a 'Girls' school are 0.7 the odds of a student attending a coeducational centre. This last result could be due to a school selection and/or motivation effect: highly motivated girls who wanted to study Chemistry might have decided to attend a mixed sixth form because of the traditional belief that 'Boys schools' were better at science subjects (many 'Boys schools' have mixed sixth forms).

The size of the school does not seem to be associated with the students' success in Chemistry A-level.

Although many of the effects of the individual and school characteristics can be understood and interpreted by observing the coefficients in previous tables, it is always useful to consider a plot of these effects (Figure 2). Any variable whose line intersects with the vertical zero axis can be regarded as not significant (at the 5% level) and the length of the line gives an indication of the relative size of the group, for example, the number of Caribbean students is low. Positive values imply a positive relationship with the outcome; negative values imply that the probability of obtaining grade A in Chemistry at A-level decreases with higher values of the background variable. From this figure, we can see at a glance which variables are strongly related to the probability of obtaining grade A, both positively and negatively, and which ones seem to have much less definite relationships, even if they are statistically significant.

The variable that has the largest positive effect on obtaining grade A in Chemistry A-level is the prior attainment at GCSE. The average performance of the students in a centre (school attainment) is also a significant predictor of individual success at A-level. The effects of centre type are small in comparison, in particular the effects of attending a grammar or an independent school.

Based on this graph, although the prior attainment has the highest impact on the probability of obtaining grade A at A-level, other factors such as school characteristics explain a substantial proportion of the

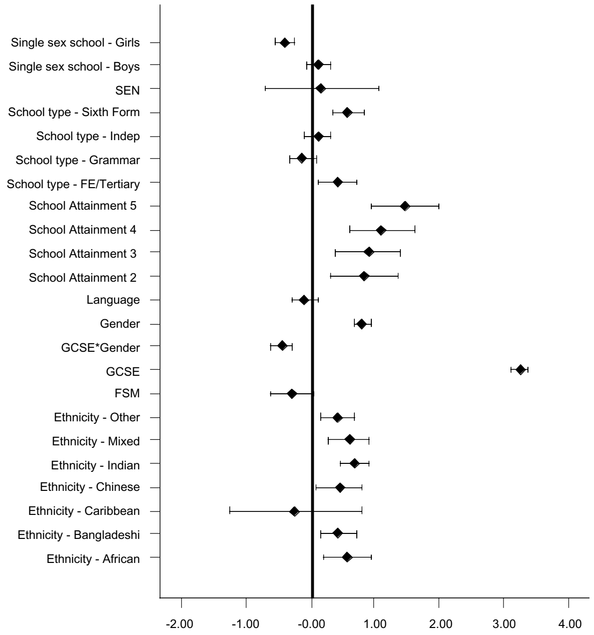


Figure 2 : Effects of the individual and school characteristics (grade A)

variation in the students' outcomes. For the model in Table 6, the explained proportion of the variance was computed (Snijders and Bosker, 1999) and it has a value of 0.59. The unexplained proportion can be partitioned between school and candidate as 0.05+0.36, which means that 5% of the variation is unexplained variation at the school level and 36% is unexplained variation at student level.

Further models that included socio-economic factors in the form of neighbourhood characteristics were fitted. Their effects, without taking into account prior attainment, are shown in Figure 3.

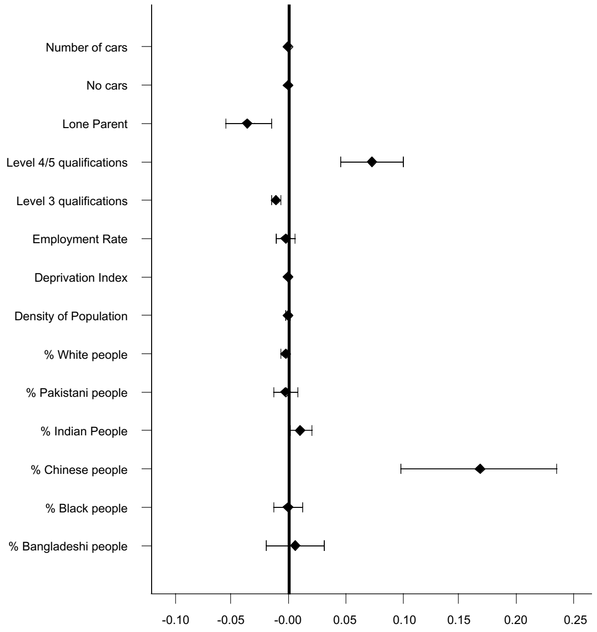


Figure 3 : Effects of the neighbourhood characteristics – no prior attainment (grade A)

The variable that has the largest positive effect on obtaining grade A is being in a neighbourhood with a high percentage of Chinese people. Also, a school situated in a neighbourhood where there is a large percentage of people with at least Level 4/5 qualifications has a positive effect on the students' success in Chemistry A-level. An area with a high number of lone parents with children has an important and significant effect but in this case negative. The impact of these factors is very small in comparison with the effect of the prior attainment or some of the school characteristics (see Figure 2).

Table 7 shows only the significant neighbourhood characteristics after adjusting for prior attainment. Their distributions can be found in Vidal Rodeiro (2006). Prior attainment measures are likely to incorporate deprivation effects that operate during earlier childhood years, and we should therefore be conservative in our estimates of the magnitude of the total effect of deprivation.

Table 7 : Neighbourhood characteristics (prior attainment)

	Estimate	*Standard Error*	*Odds Ratio*
Lone parent	**-0.03506**	0.01034	**0.9**
Level 3 qualifications	**-0.01029**	0.00234	**0.9**
Level 4/5 qualifications	**0.07281**	0.01389	**1.1**
% Chinese people	**0.16751**	0.03504	**1.2**
% Indian people	**0.01102**	0.00487	**1.0**
Deprivation Index	**-0.00039**	0.00018	**0.9**

Grade E

Do the factors (individual, school or neighbourhood characteristics) that have an effect on the probability of obtaining grade A, have an effect on the probability of obtaining other grades? In this section, we repeat the previous analyses but the dependent variable takes the value 1 if the student obtains at least grade E and 0 otherwise.

For at least grade E, gender and mean GCSE are still statistically significant. However, the odds of obtaining at least grade E for a boy are only 1.3 times the odds of a girl obtaining at least grade E compared to the 2.2 for grade A. The effect of prior attainment is, as expected, much lower.

Having a mixed background or being Chinese does not have a significant effect on the probability of obtaining at least grade E. Being Bangladeshi, African, Indian or part of other ethnic groups has a positive significant effect on the outcome and this effect is larger for at least grade E than for grade A.

Another difference between grade A and at least grade E is that for the latter the first language has a significant effect and the odds of obtaining at least grade E for a student with a first language other than English are 0.6 times the odds of a student whose first language is English.

In a following step a model with the school level characteristics plus students' prior attainment was fitted. The odds of obtaining at least grade E for a student attending a particular type of school are very similar to those for grade A. However, only 'attending a FE/Tertiary college' has a significant effect on the outcome. As before, the school attainment group plays an important role in the success of a student taking Chemistry A-level. The higher the attainment of the school, the larger the odds of obtaining at least grade E. The effect of school gender and school size is the same as for grade A.

To summarise, Figure 4 displays the effects of the individual and school type characteristics on the probability of obtaining at least grade E.

Table 8 shows the results obtained when models were fitted with socio-economic factors taken into account. The factor that has the largest positive significant effect on the probability of obtaining at least

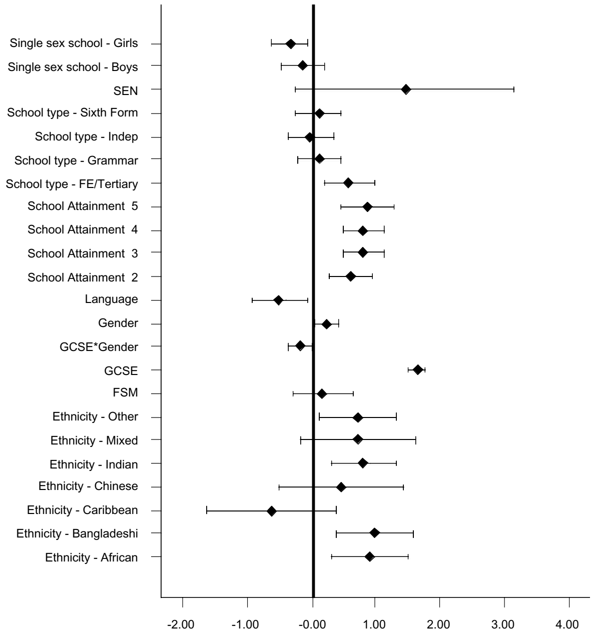


Figure 4 : Effects of the individual and school characteristics (at least grade E)

grade E is the employment rate. As the percentage of employed people in the neighbourhood increases, so the probability of obtaining at least grade E does. Being in a neighbourhood with high percentages of single-parent families has a negative effect on the probability. Another factor that has a negative effect is the percentage of ethnic minorities (Chinese, Indian, Black and Bangladeshi). Being in a neighbourhood with high percentages of white people has a positive effect on attainment.

Table 8 : Neighbourhood characteristics (prior attainment)

	Estimate	*Standard Error*	*Odds Ratio*
Level 3 qualifications	**0.01782**	0.00374	**1.0**
Level 4/5 qualifications	**-0.07089**	0.02276	**0.9**
% Chinese people	**-0.27535**	1.50448	**0.7**
% Black people	**-0.03296**	0.00808	**0.9**
% White people	**0.01170**	0.00257	**1.0**
% Indian people	**-0.01925**	0.00610	**0.9**
% Bangladeshi people	**-0.03499**	0.01311	**0.9**
No cars	**-0.00011**	0.00004	**1.0**
Employment Rate	**0.02433**	0.00608	**1.0**
Population Density	**-0.00794**	0.00134	**0.9**

Conclusions and discussion

The effects of basic explanatory factors (e.g. prior attainment and gender) made statistically significant contributions to the success in A-level Chemistry. Having taken into account prior attainment, several school effects also proved significant, in particular the average performance of the students in a school is a significant predictor of individual success.

With regard to the effect of the school type, given a mean GCSE score, the probability of obtaining grade A is slightly higher if the student attended a sixth form college or an FE/Tertiary college than if the student attended a grammar or a comprehensive school. However, if a sixth form has an able entry, that is, has many pupils who did very well in their GCSEs, then, on average, it does not matter which type of school it is. However, these are average effects and there is still considerable variation between individual schools that is large enough to cancel out these effects.

Substantial differences appeared between ethnic groups. The results show that in comparison to the white group, other ethnic groups have significantly higher probability of obtaining grade A but their effects are not significant when modelling the probability of obtaining at least grade E. However, all the differences described might not be attributed entirely to ethnicity. Different ethnic groups have different socio-economic profiles and consequently it is not possible to say categorically whether the differences observed are the result of ethnic differences per se or whether socio-economic or other factors play a part.

By comparing the significant explanatory variables included in the different models, our findings show that when prior attainment data are lacking, other student background and school context information explain the students' success at A-level Chemistry. However, prior attainment has, by far, the largest impact on the success.

All models give very similar percentages of the school and pupil level variance explained, but the one described in Table 6 gives the highest percentages, showing that school characteristics (type, attainment and gender) explain more about the students' performance than the neighbourhood characteristics. The unexplained percentage of the variation in the models fitted in this article is around 42%. The amount of school level variance unexplained is relatively small (around 4–5%) but the unexplained variation at student level is around 36–38%, suggesting that the individual students' characteristics are much more important than the school they attend. Also, the amount of unexplained variation at student level could be due to the fact that other variables that have not been included in the model (e.g. subject preference, motivation) may have an influence on students' success.

A conceptual limitation of all regression techniques is that one can only ascertain relationships, but never be sure about the underlying causal mechanism. Therefore, caution must be taken when interpreting the results of the regression analyses shown in this article. In this research, we found significant relationships between some individual, school or socio-economic characteristics and attainment. However, they may not be the result of a causal relationship.

References

Bell, J.F. (2003). Beyond the school gates: the influence of school neighbourhood on the relative progress of pupils. *Oxford Review of Education*, **29**, 4, 485–502.

Ensminger, M.E., Lankin, R.P. & Jacobson, N. (1996). School leaving: a longitudinal perspective including neighbourhood effects. *Child Development*, **67**, 5, 2400–2416.

Gray J., Jesson, D. & Sime, N. (1990). Estimating differences in examination performance of secondary schools in six LEAs: a multilevel approach to school effectiveness. *Oxford Review of Education*, **16**, 2, 137–158.

Haque, Z. & Bell, J.F. (2001). Evaluating the performances of minority ethnic pupils in secondary schools. *Oxford Review of Education*, **27**, 3, 357–368.

Malacova, E. (2006). Effect of single-sex education on progress in GCSE. *Oxford Review of Education*. In press.

OECD (2004). *Learning for Tomorrow's World. First Results from PISA 2003*. Paris: OECD Publications.

Raffe, D., Croxford, L., Iannelli, C., Shapira, M. & Howieson, C. (2006). *Social-Class Inequalities in Education in England and Scotland*. Special CES Briefing No. 40. Edinburgh: CES.

Rasbash, J., Browne, W., Healy, M., Cameron, B. & Charlton, C. (2005). *MLwiN version 2.02*. London: Institute of Education.

Rutter, M.L., Maugham, B., Mortimore, P., Ousten, J. & Smith, A. (1979). *Fifteen thousand hours*. London: Open Books.

Snijders, T. and Bosker, R (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Vidal Rodeiro, C.L. (2006). *Factors determining examination success at A-level: a study focussed on A-level Chemistry and A-level Business Studies*. Internal Report. Cambridge: Cambridge Assessment.

A-level uptake: 'Crunchier subjects' and the 'Cracker effect'

John F. Bell, Eva Malacova, Carmen L. Vidal Rodeiro Research Division
Mark Shannon New Developments

One of the claims made about A-levels is that students are opting for the allegedly easier subjects at A-level. For example, Boris Johnson stated in the *Observer* (July 9, 2006) that 'This year, as every year for the last two decades, we are seeing a drift away from crunchier subjects such as sciences, maths and languages.' More recently, Cambridge University produced a list of A-level subjects that provide a less effective preparation for their courses, for example, Business Studies, Media Studies, and Physical Education, Sports Studies. On their website (http://www.cam.ac.uk/admissions/undergraduate/requirements/), it is stated 'To be a realistic applicant, Cambridge applicants would be expected to have no more than one of these subjects'[1]. It must be stressed that the term 'less effective preparation' refers to the courses offered by what is a highly selective university – these A-levels can be highly relevant and effective preparations for courses offered by other higher education institutions. It is also worth noting that some subjects not on the list had to struggle to gain acceptance. For example, Tillyard (1958) wrote:

> ... *[in 1878] it was unthinkable that English should be recognised as an independent study; it could enter Cambridge only on the warrant of a faint respectability reflected from modern languages.*

Opponents of English could be quite outspoken, for instance, Edward Augustus Freeman, the Regius Professor of Modern History at Oxford, in a broadside published in 1887 in the London *Times* wrote:

> *There are many things fit for a man's personal study, which are not fit for University examinations. One of these is "literature."... [We are told] that it "cultivates the taste, educates the sympathies, enlarges the mind." Excellent results against which no one has a word to say. Only we cannot examine in tastes and sympathies.*

As late as 1965, Robson used the first lecture arranged by the F.R. Leavis Lectureship Trust to argue that English Studies met the conventional criteria for admission to a *studium generale*[2]. Also, in 1887 the congregation of Oxford University voted against an Honour School of Modern European Languages. The Warden of All Souls objected because of 'the depreciation and exclusion of Greek and Latin' and that 'it confused the whole conception of academical studies, and dragged the subjects fit for more advanced years into undergraduate life' (*Times*, 7 November, 1887). However, as Emperor Loathair I (795–855) said 'Tempora mutantur, nos et mutamur in illis'.[3] Whilst it might be possible to idly speculate what a Regius Professor of Media Studies at a 22nd century Cambridge University might make of the current situation, it is probably more informative to consider what exactly is happening with A-levels and determine if the changes are as dramatic as is implied in the media.

In this article we investigate the uptake of A-levels in England. We consider the A-level results for all year 13 students (eighteen-year-olds) in 2001 to 2005 (more detailed analyses for the earlier years can be found in Bell, Malacova and Shannon, 2003, 2005[4]). This period covers the transition to Curriculum 2000 because the new A-levels that were started then were completed in 2002. This reform split A-levels into two. First, a free standing qualification called the Advanced Subsidiary covering the first year of the course was introduced. Secondly, the A-level was obtained by combining results of AS modules with A2 modules. The aim of this reform was that students would study for four or five subjects at AS in the first year of the sixth form and then choose three of them to continue on to A-level. The objective of this reform was to broaden the curriculum and to provide more balance. This is seen as a desirable outcome in many areas of higher education. For example, all medical schools (except Dundee and Edinburgh) encourage potential applicants to take a combination of science and non-science subjects (Clarke, 2005). These medical school policies have implications for the A-level science uptake for the higher attaining candidates. Given that Chemistry is nearly always compulsory and Biology often is, then the effect would be most pronounced in Physics and Mathematics. In general, there are two processes that need to be considered. First, broadening the

[1] There are exceptions and it is always advisable to check the Cambridge University website for the precise requirement for a course.

[2] A recognised university. Originally an institution recognised by the Holy Roman Empire and whose status was confirmed by Papal Bull. Cambridge was formally acknowledged as one in 1290.

[3] For those who have not had a classical education: 'Times are a-changing and we change with them.'

[4] The analyses differ from those in this article because they include General Studies A-level.

curriculum would lead to a decline in the number of science A-levels as science specialists are encouraged to take other subjects. Secondly, the reverse process is true for non-scientists. In both cases, the change is likely to affect the student's least favourite or least relevant subject. This argument means these changes only affect the pool of qualified candidates actually applying for higher education courses when students change their future plans during their A-level studies, for example, students who would have taken only science A-levels but who substituted Physics with a non-science would have been unlikely to continue with Physics after A-level had they remained science specialists.

Uptake of the most common A-level subjects

In Tables 1–3, the uptake of A-level subjects (strictly, the subject has been defined by the subject code used in the database rather than the specification name) with the highest entries is presented. For clarity, these subjects have been divided into three groups: science and mathematics; arts, languages and sports science; social science and humanities. Candidates were classified by sex and prior attainment at GCSE level. The GCSE grades for the candidates were converted into scores (A*–8, A–7, B–6, etc.) and a mean GCSE score was computed and used to divide the candidates into three attainment groups: low, medium and high. The cut scores were chosen such that they divided the whole A-level entry in three approximately equally sized groups and were carried over for future years. Uptake by attainment is an important issue. Since elite institutions are more likely to require good grades and candidates with higher prior attainment are more likely to obtain such grades, it follows that changes in uptake of subjects by high attaining students can have important implications on the pool of available applicants for courses at these institutions.

Table 1 presents the percentages of A-level students (i.e. having one A-level result) taking each of English and modern language subjects. Changes over the whole period greater than 2 percentage points have been identified in **bold** for declines and *italics* for increases. For all the subjects in this group, uptake is much greater for female students. Uptake also increases with increasing prior attainment for English Literature and the modern languages. There has been a decline in the uptake by female students for English Literature, French and German. This decline is also associated with medium and high prior attainment.

In Table 2 uptakes for arts, media studies and sport studies are presented. The highlighted trends are a decline of females taking Art and Design associated with the change to Curriculum 2000, an increase in Drama uptake, an increase in Media Studies at the time of the change to Curriculum 2000 and an increase in uptake of Sports Studies by candidates with medium prior attainment.

Table 3 is for the humanities and social sciences. Three subjects, Business Studies, Economics, and Geography are all declining except for candidates with high prior attainment. The 'crunchy' subject, History, had an increased uptake by males as did Politics. The increase in Politics was associated with male and high prior attainment candidates. Religious studies was also increasing in popularity. However, the largest changes are associated with Psychology which has the largest increase of any A-level subject, for example, almost one in four females taking three or more A-levels take Psychology. This has been referred to as the 'Cracker[5] effect' because it is argued that uptake has been influenced by the

Table 1 : Changes in uptake in English and Modern Languages

(% of students with at least one A-level result)

	Year	*English*	*English Language*	*English Literature*	*French*	*German*	*Spanish*
All	2001	7	6	21	7	**4**	2
	2002	6	7	21	6	**3**	2
	2003	6	6	20	6	**3**	2
	2004	6	6	20	6	**2**	2
	2005	6	8	20	6	**2**	2
Male	2001	4	4	13	4	2	1
	2002	4	5	14	4	2	1
	2003	4	5	12	4	2	1
	2004	4	4	12	4	2	2
	2005	4	6	14	4	2	2
Female	2001	9	8	**29**	**9**	**4**	3
	2002	8	8	**27**	**8**	**4**	3
	2003	7	8	**26**	**7**	**3**	3
	2004	8	8	**26**	**6**	**4**	2
	2005	8	8	**26**	**6**	**2**	2
Low	2001	8	7	16	2	1	1
	2002	7	7	17	1	1	1
	2003	6	7	15	1	1	1
	2004	6	6	14	2	2	0
	2005	6	8	16	2	2	0
Medium	2001	7	7	**23**	**6**	3	2
	2002	7	8	**21**	**4**	2	2
	2003	7	8	**20**	**4**	2	1
	2004	6	8	**20**	**4**	2	2
	2005	8	8	**20**	**4**	2	2
High	2001	5	4	**26**	**14**	**6**	4
	2002	6	5	**25**	**12**	**5**	4
	2003	5	5	**24**	**11**	**5**	4
	2004	4	6	**22**	**10**	**4**	4
	2005	4	6	**24**	**10**	**4**	4

Table 2 : Changes in uptake of Arts, Media and Sport/PE studies

(% of students with at least one A-level result)

	Year	*Art & Des.*	*Drama*	*Media/Film/ TV. Stds.*	*Music*	*Sport/ P.E.Stds.*
All	2001	9	*4*	*6*	2	7
	2002	10	*6*	*8*	3	7
	2003	10	*6*	*8*	3	8
	2004	8	*6*	*8*	2	8
	2005	8	*6*	*8*	2	8
Male	2001	7	*2*	*5*	2	9
	2002	7	*3*	*7*	2	10
	2003	8	*4*	*7*	2	11
	2004	8	*4*	*8*	2	10
	2005	8	*4*	*8*	2	10
Female	2001	**11**	*5*	*6*	3	5
	2002	**8**	*8*	*8*	3	5
	2003	**9**	*8*	*9*	3	5
	2004	**8**	*8*	*10*	2	6
	2005	**8**	*8*	*10*	2	6
Low	2001	10	*4*	*9*	1	10
	2002	11	*7*	*12*	2	10
	2003	11	*7*	*13*	1	11
	2004	10	*6*	*14*	2	10
	2005	10	*6*	*14*	2	10
Medium	2001	10	*4*	*6*	2	*8*
	2002	11	*7*	*9*	2	*9*
	2003	12	*7*	*9*	2	*10*
	2004	10	*6*	*10*	2	*10*
	2005	10	*8*	*10*	2	*10*
High	2001	8	*3*	*2*	4	*3*
	2002	9	*5*	*3*	3	*4*
	2003	10	*5*	*4*	4	*5*
	2004	10	*6*	*4*	4	*4*
	2005	10	*6*	*4*	4	*6*

[5] A slang term for a criminal psychologist and used as a name for a popular TV series about one.

Table 3 : Changes in uptake in Humanities and Social Sciences

(% of students with at least one A-level result)

Group	*Year*	*Busi. St.*	*Econ.*	*Geog*	*History*	*Law*	*Politics*	*Psych.*	*Relig. Stds.*	*Socio.*
All	2001	**14**	7	**15**	*15*	*3*	3	*10*	*3*	9
	2002	**13**	6	**15**	*17*	*4*	3	*13*	*4*	9
	2003	**13**	6	**14**	*16*	*4*	4	*14*	*5*	9
	2004	**12**	6	**14**	*16*	*4*	4	*16*	*6*	10
	2005	**12**	6	**12**	*18*	*6*	4	*18*	*6*	10
Male	2001	**16**	**10**	**18**	*15*	3	*4*	*5*	*2*	4
	2002	**17**	**9**	**17**	*17*	3	*4*	*6*	*2*	4
	2003	**16**	**9**	**17**	*17*	4	*5*	*7*	*3*	4
	2004	**16**	**8**	**16**	*18*	4	*4*	*8*	*4*	4
	2005	**14**	**8**	**14**	*18*	4	*6*	*10*	*4*	4
Female	2001	**12**	**4**	**13**	15	*4*	3	*15*	*5*	13
	2002	**10**	**3**	**12**	16	*4*	3	*18*	*5*	13
	2003	**10**	**3**	**12**	16	*5*	3	*20*	*6*	13
	2004	**8**	**2**	**12**	16	*6*	2	*22*	*6*	12
	2005	**8**	**2**	**10**	16	*6*	2	*24*	*8*	14
Low	2001	**17**	**4**	**11**	10	5	2	*11*	3	13
	2002	**15**	**3**	**11**	11	5	2	*12*	3	12
	2003	**15**	**2**	**10**	10	5	2	*13*	4	12
	2004	**14**	**2**	**8**	10	6	2	*14*	4	12
	2005	**12**	**2**	**8**	10	6	2	*14*	4	12
Medium	2001	**17**	**7**	**18**	16	*3*	3	*12*	*4*	*10*
	2002	**16**	**5**	**16**	16	*4*	3	*16*	*4*	*11*
	2003	**16**	**5**	**16**	16	*5*	3	*18*	*5*	*11*
	2004	**14**	**4**	**14**	16	*6*	4	*20*	*6*	*12*
	2005	**14**	**4**	**14**	16	*6*	4	*22*	*6*	*12*
High	2001	8	9	17	21	*2*	*4*	*7*	*4*	*4*
	2002	9	9	17	22	*2*	*4*	*11*	*5*	*5*
	2003	8	9	16	22	*3*	*5*	*13*	*5*	*5*
	2004	8	8	16	22	*4*	*4*	*14*	*6*	*6*
	2005	8	8	16	22	*4*	*6*	*14*	*6*	*6*

Table 4 : Uptake of Science and Mathematics subjects

(% of students with at least one A-level result)

	Year	*Biology*	*Chem.*	*Com. Stds*	*D & T design*	*ICT*	*Maths*	*Further Maths*	*Physics*
All	2001	19	16	**5**	*2*	*3*	**24**	2	**13**
	2002	19	14	**4**	*6*	*7*	**19**	2	**13**
	2003	18	13	**4**	*6*	*7*	**19**	2	**12**
	2004	18	14	**2**	*6*	*6*	**20**	2	**10**
	2005	18	14	**2**	*6*	*6*	**18**	2	**10**
Male	2001	16	18	**9**	*4*	*4*	**32**	4	**22**
	2002	16	15	**8**	*9*	*10*	**26**	3	**21**
	2003	15	14	**7**	*9*	*10*	**26**	3	**20**
	2004	14	14	**6**	*9*	*8*	**26**	4	**18**
	2005	16	16	**4**	*9*	*8*	**26**	4	**18**
Female	2001	**22**	**14**	1	*1*	*2*	**17**	1	5
	2002	**22**	**14**	1	*4*	*4*	**13**	1	5
	2003	**20**	**13**	1	*4*	*5*	**13**	1	5
	2004	**20**	**12**	*	*4*	*4*	**14**	*	4
	2005	**20**	**12**	*	*4*	*4*	**12**	*	4
Low	2001	*9*	5	6	3	4	8	*	5
	2002	*7*	3	4	8	9	4	*	4
	2003	*6*	3	3	8	10	4	*	3
	2004	*6*	4	2	8	8	6	*	4
	2005	*6*	4	2	8	8	6	*	4
Medium	2001	**19**	**12**	**6**	*3*	*3*	**21**	1	**11**
	2002	**17**	**10**	**5**	*7*	*8*	**13**	1	**10**
	2003	**16**	**8**	**4**	*8*	*9*	**13**	1	**9**
	2004	**14**	**8**	**4**	*8*	*8*	**12**	*	**8**
	2005	**14**	**8**	**2**	*8*	*6*	**12**	*	**8**
High	2001	30	**30**	3	**1**	1	**42**	**6**	**22**
	2002	30	**27**	3	**4**	4	**35**	**4**	**21**
	2003	29	**25**	3	**4**	4	**34**	**4**	**19**
	2004	28	**26**	2	**4**	4	**34**	**4**	**18**
	2005	30	**26**	2	**4**	2	**32**	**4**	**16**

* denotes less than 0.5% uptake

increasing prominence of psychologists in television drama. (http://news.bbc.co.uk/1/hi/education/1635122.stm)

Finally, we consider the uptake of science and mathematics subjects. The longer term trends in Mathematics were considered in Issue 2 of *Research Matters* (Bell, 2006). There have been declines in uptake for females taking Biology and Chemistry. These are particularly associated with medium levels of prior attainment. For Physics the decline also occurs for high attaining candidates. It is important to note that throughout the period under consideration entries in the three traditional sciences and Mathematics have been dominated by candidates with high levels of GCSE attainment. This raises the question as to whether it is desirable for advanced studies in these subjects to be increasingly the preserve of an academic elite. There are clearly issues with the perceived difficulty of these subjects.

It is not enough to consider uptake of individual subjects. Combinations of subjects are also important as indicated by the University of Cambridge's concerns mentioned in the first paragraph of this article. Analysing combinations of subjects is not as straightforward as it seems because there were 23,963 combinations of individual subjects in 2001 (Bell, Malacova and Shannon, 2003, 2005). Therefore, it is necessary to group subjects to analyse combinations.

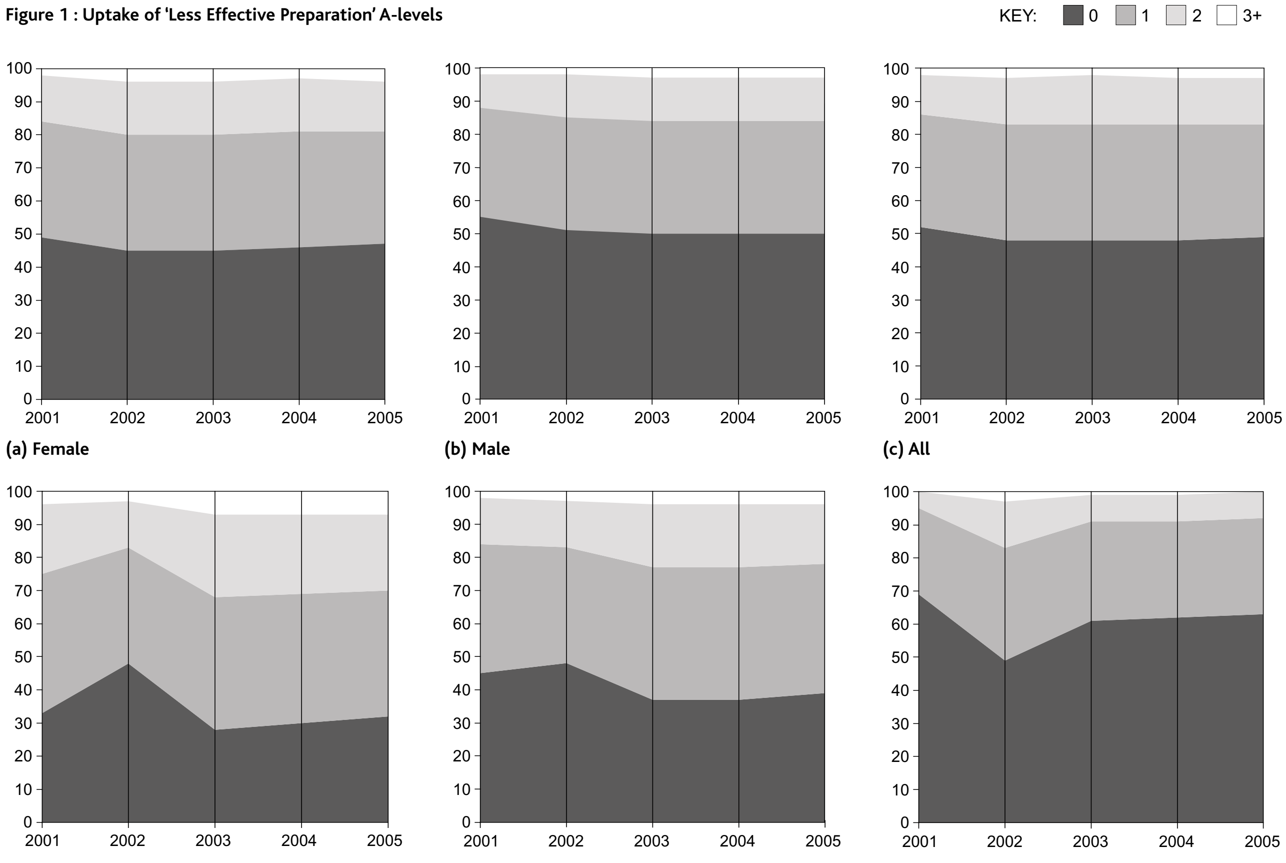
The first group to be considered is based on the Cambridge list of subjects that are less effective preparation for Cambridge courses (LEPs) (http://www.cam.ac.uk/admissions/undergraduate/requirements/). The subjects that they list are: Accounting, Art and Design, Business Studies, Communication Studies, Dance, Design and Technology, Drama/Theatre Studies, Film Studies, Health and Social Care, Home Economics, Information and Communication Technology, Leisure Studies, Media Studies, Music Technology, Performance Studies, Performing Arts, Photography, Physical Education, Sports Studies, and Travel and Tourism.

It is not the study of individual LEPs that is the perceived problem but rather the studying of too many of them. We decided to investigate the number of these LEPs taken by candidates with three or more A-level results. In Figure 1, 100% stacked area charts are presented for the number of less effective preparation A-levels (LEPs). The data is presented by gender and by prior attainment. The prior attainment is based on mean GCSE with the cut scores that divided the 2001 A-level candidates into three approximately equal groups. The darkest area at the bottom represents the candidates not taking any LEPs. The lighter grey area represents those taking one LEP and the next two areas 2 and 3+ LEPs (3+ is the top area). The data for all the tables in this report can be found in Vidal Rodeiro (2006) on the Cambridge Assessment website. http://www.cambridgeassessment.org.uk/research/statisticalreports/

It is clear that the majority of students still take at least two 'acceptable' subjects. Male students are less likely to take LEPs than female students. There is an interesting effect relating to Curriculum 2000. In the first year, there was a decrease in LEP subjects by low prior attainment candidates and the reverse pattern for high prior attainment candidates. It is likely that pattern was based on centres gaining experience of the new specification (in particular the A2 modules) and the attitudes of universities to certain subjects.

It should be recognised that the majority of candidates for Cambridge and other elite institutions likely to have similar restrictions on suitability of A-levels will largely recruit from candidates with high prior GCSE attainment. Only 5% of the candidates in 2001 made a choice of A-levels that included 2 or more LEPs. This increased to 17% in 2002 and then fell to about 9% for the remaining three years (note that for candidates performing at the level of the vast majority of successful Cambridge applicants, i.e. those with three grade As, the percentages affected are much smaller.)

Figure 1 : Uptake of 'Less Effective Preparation' A-levels



(a) Female

(b) Male

(c) All

(d) Low

(e) Medium

(f) High

One of the aims of Curriculum 2000 was to broaden students' experiences and to discourage early specialisation. In Figure 2, the uptake of five subject areas at A-level is presented. The five areas were: Science/Mathematics, English, Languages, Social Science/Humanities and Arts. Grouping subjects is not a straightforward task and the allocation of subject areas is always debatable (at the time of analysis psychology specifications were usually grouped with the social sciences but modified specifications starting in 2008 are going to be classified as sciences). Some subjects do not necessarily fit comfortably in any category. More details of the subject areas can be found in Bell, Shannon and Malacova. (2003, 2005). The categories were originally derived to illustrate how close the current situation is to a balanced diploma based on existing A-levels. The percentages in the figure relate to the number of A-level students taking at least one of the subjects in the subject area in the population of students taking at least three A-levels.

The most obvious feature of Figure 2 is the stability of the uptake for most subject areas. The clearest trend is the decline in the number of students taking Modern Languages where the rate of decline is faster for female students. There were no consistent trends for the other domains.

There are however large differences between the subject areas. For female students, the subject area with the highest uptake is Social Science/Humanities (note this is made up of many more A-level subjects than the other areas). The remaining subject areas in descending order of uptake are English, Science/Mathematics, Arts and Modern Languages. The pattern for male students is different. The uptake of the Science/Mathematics subject area is similar to that of the Social Science/Humanities area. The uptakes of subjects in the English, Arts and the Modern Languages groups are much lower than for females.

Figure 2 also shows that the uptake of subject domains is related to ability. Uptake of Arts, English, and Social Science domains all decline with increasing prior attainment. This is most marked for the Arts domain where the percentage uptake is approximately halved. Uptake for the remaining two domains, Modern Languages and Sciences, increases with prior attainment. This relationship is strongest for Modern Languages with approximately one in twenty students in the lowest prior attainment group taking at least one modern language compared with one in five for the high prior attainment group.

John Dunford, general secretary of the Association of School and College Leaders stated that Modern Languages were in freefall (*Guardian*, 24 August, 2006). Whilst the decline has been substantial, there are two features of it that are interesting. First, it is much smaller for males compared with females and, secondly, for low attaining candidates the decline is also smaller. There is a need for further research in these areas.

For the final analyses in this article, subjects were grouped into three different domains; Science and Mathematics, Arts and Languages, Social Science and Humanities (the subject domains involved merging the subject areas of Art, English and Languages into one domain. This categorisation was used in Bell *et al.* (2005) to investigate whether A-level subject choice was balanced). Using these three domains, it possible to classify candidates taking three or more A-levels into seven groups:

YNN Science/Mathematics only.

NYN Arts /Languages only

NNY Social Sciences/Humanities only

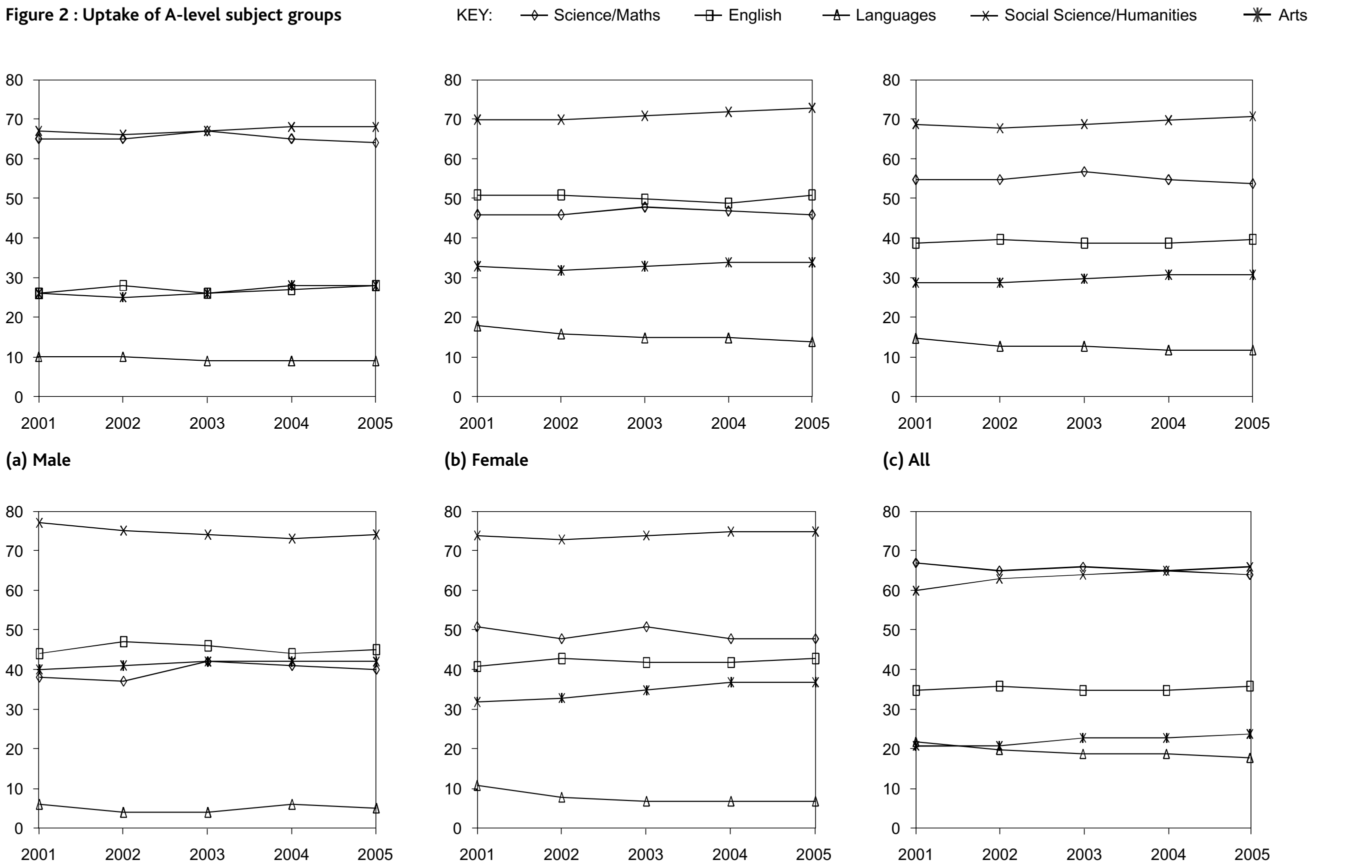
YYN Science/Mathematics and Arts/Languages

YNY Science/Mathematics and Social Sciences/Humanities

NYY Arts / Languages and Social Sciences / Humanities

YYY All three domains

Figure 2 : Uptake of A-level subject groups



(a) Male

(b) Female

(c) All

(d) Low

(e) Medium

(f) High

Figure 3 presents the percentage uptake of these categories by sex and attainment for 2001 to 2005 for those candidates taking at least three A-levels. In the figures solid lines have been used to denote combinations including science and mathematics and dotted lines for those that do not. Looking at all the panels of Figure 3 it can be seen that for most combinations of domains the changes have been small. The only large changes tend to be associated with science specialists. The increases tend to be smaller and shared between combinations involving science. For the combinations of domains not including Science/Mathematics, there are only small, inconsistent year on year changes. This means that the net effect of broadening the curriculum has been to reduce the amount of science that science specialists study but this has not been matched by an increased uptake by non-scientists.

When all the data presented here are considered in their entirety, there are some noticeable results. First, for most subjects and groups of subjects there has been very little change. For some subjects and groups of subjects there have been changes associated with Curriculum 2000 but the uptakes have subsequently stabilised. Of greater concern are the subjects that have declined through the whole period, for example, Geography, Physics and Modern Languages as a group.

Although there has been a decline in numbers taking general qualifications (GCSE and GCE) that assess modern languages, there is an important development that seems to provide a promising solution. This summer's GCSE results showed a big decline in the number of pupils studying modern foreign languages. French and German suffered the biggest falls in candidates of any subject, with declines of 13.2% and 14.2%. There are, however, alternatives to existing qualifications that may be of use in increasing the number of linguists. Cambridge Assessment has developed a new qualification scheme called Asset Languages. This is part of the DFES National Languages Strategy (http://www.assetlanguages.org.uk/). They use a 'ladder' of courses similar to music grades and aim to make language learning accessible. More than a quarter of state secondary schools are going to use these qualifications from September 2006. There are also 120 primary schools involved in the scheme. Experience from the first full year of the scheme suggests that it is successful in motivating students' language learning.

For Science and Mathematics, there is a need to consider how these subjects are extended beyond a very able elite. When considering trends in uptake, a common mistake is to use what was described in the TV series *Yes Minister* as the politician's syllogism: 'Something must be done. This is something. Therefore, this must be done'. Before acting it is better to gain an understanding of the underlying causes for the trend. This article is only the first step in understanding uptake in A-levels. Before acting it is necessary to understand the processes that have led to the situation described in this article. This requires the collection of additional information. For this reason, Cambridge Assessment is currently conducting a large scale survey (with the Association of Colleges) investigating why students choose particular A-levels.

Acknowledgements

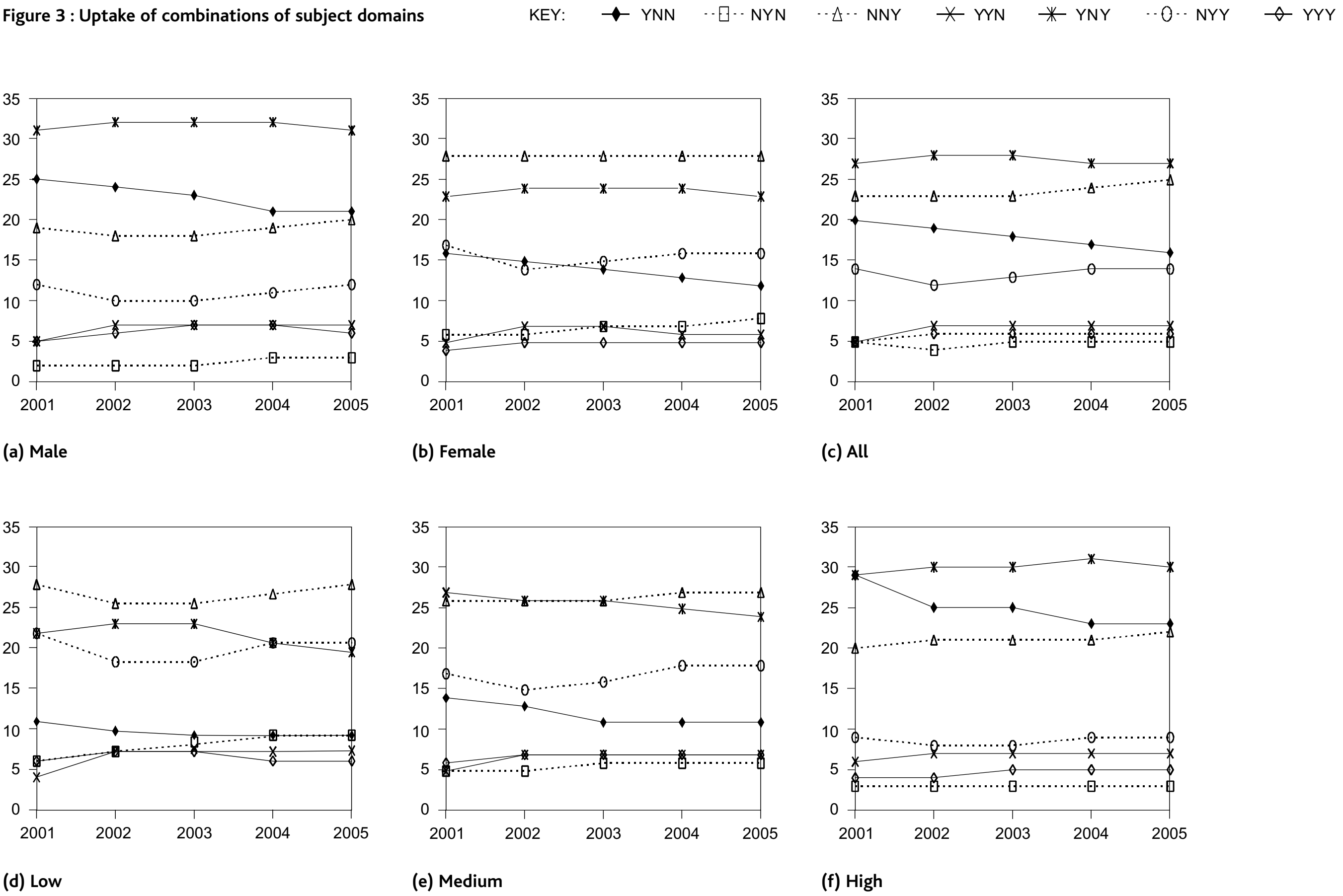
The authors would like to thank the Cambridge Assessment Librarian, Fiona Beedle, for her assistance in tracking down the older references in this article.

References

Bell, J.F. (2006). The curious case of the disappearing mathematicians. *Research Matters: A Cambridge Assessment Publication*, **2**, 21–23.

Bell, J.F., Malacova, E. & Shannon, M. (2003). *The Changing Pattern of A-level/AS uptake in England*. A paper presented at the British Educational Research

Figure 3 : Uptake of combinations of subject domains



(a) Male

(b) Female

(c) All

(d) Low

(e) Medium

(f) High

Association Annual Conference, Edinburgh, September 2003. http://www.cambridgeassessment.org.uk/research/confproceedingsetc/BERA2003JBEMMS/file/

Bell, J.F., Malacova, E. & Shannon, M. (2005). The changing pattern of A level/AS uptake in England. *The Curriculum Journal*, **16**, 3, 391–400.

Clarke, P. (2005). *Curriculum 2000 and other qualifications. A survey of UK medical schools' requirements and selection policies*. Cardiff University. http://www.ucas.ac.uk/candq/curr2000/medical.pdf

Centre for Education and Industry (2003). *Widening participation in the physical sciences: An investigation into factor influencing the uptake of physics and chemistry*. University of Warwick, Coventry: Centre for Education and Industry.

Robson, W. W. (1965). *English as a university subject*. London: Cambridge University Press.

Smithers, A. & Robinson, P. (2006). *Physics in schools and universities: II. Patterns and policies*. Buckingham: Centre for Education and Employment Research. http://www.buckingham.ac.uk/education/research/ceer/pdfs/physicsprint-2.pdf

Tillyard, E.M.W. (1958). *The muse unchained: An intimate account of the revolution in English Studies at Cambridge*. London: Bowes and Bowes.

Vidal Rodeiro, C.L. (2006). Uptake of GCE A-level subjects in England 2001–2005. Statistical Report Series No. 3. http://www.cambridgeassessment.org.uk/research/statisticalreports/

PSYCHOMETRICS CENTRE

Discussion piece: The psychometric principles of assessment

Professor John Rust Psychometrics Centre

Psychometrics is the science of psychological assessment, and is a foundation of assessment and measurement. Within psychometrics there are four fundamental principles whereby the quality of an assessment is judged. These are (1) reliability, (2) validity, (3) standardisation and (4) freedom from bias. Reliability is the extent to which an assessment is free from error; validity is the extent to which a test or examination assesses what it purports to assess; standardisation gives us information on how the result of an assessment is to be judged, and freedom from bias examines the extent and causes of differences between groups. These four principles inform not only test use but also the entire process of test development, from the original curriculum or job specification, via the choice and appraisal of examination questions and test items, through to the eventual evaluation of the success or otherwise of the assessment itself.

No assessment can be perfectly reliable, and this applies not only to the measurements we make in education or psychology, but to all types of measurement. Measurements range in accuracy from the exceptionally high levels now obtained for the speed of light and the time of day, through measurements of length and area used in surveying, to the lower levels attainable for measurement of blood pressure and haematological assays used in medicine, to the tests of ability, achievement and character with which we are familiar in the education and recruitment testing world. Hence, in all these cases our expectations are different. Reliability is assessed on a scale of zero to one, with a score of 0.00 indicating no reliability at all, and a score of 1.00 representing perfect reliability. Over a century of human testing has shown us that we can expect reliabilities ranging from 0.95 for a very carefully constructed and individually assessed test of ability, through 0.85 for group tests of ability; about 0.75 for personality tests; 0.5 for rating scales and down to about 0.2 or 0.3 for projective tests or tests of creativity.

There are several ways in which reliability can be assessed and most of them involve making multiple measurements. Inter-rater reliability is the extent to which examiners agree or disagree about the mark that a candidate should be given when the assessments are made independently. This is the most relevant form of reliability for many forms of school examinations, and the assessment of coursework or essays has an apparent upper limit of about 0.65. It is interesting that in spite of frequent attempts to improve on inter-relater reliability of examiners, for example by setting detailed marking criteria, it is unusual to find a value that goes much above this figure. Constraining the judgement of individual examiners can, if taken to extremes, lead to just another multiple choice test by another name.

Whatever efforts are put into improving the agreement between markers, it is not the only source of unreliability and may not even be the most important. In multiple choice examinations, for example, the inter-rater reliability is often as high as 0.99, because the only disagreement between raters is in reading the scores from the answer sheets. However, unreliability still arises for many other reasons, such as the state of the candidate (tired, ill, anxious etc.) the environment in which the test is taken, events at home or among the peer group, or the concordance between the content of the examination and the revision strategy used by the candidate, to name but a few. These forms of reliability are called test stability, and one way of obtaining this is by administering the same test or examination to the same group of individuals on two or more occasions and comparing the results. But this can only be an estimate, as the prior experience of having sat the same examination will tend to affect the second sitting in some way. In spite of this it is essential that we have some way of estimating stability effects for our assessments. Given all the possible sources of instability, we expect an upper limit of at most about 0.85 on the expected reliability of a multiple choice school examination. It is important to note I am not here trying to do full justice to issues of reliability, I am illustrating the importance of the application of psychometric principles.

Reliability in assessment is just the first step, however. A score can be perfectly reliable and still utterly invalid for a particular application. Astrological charts or diagnoses made on the basis of graphology (handwriting analysis) may be very reliable in that astrologers or graphologists using the same system will usually be in agreement about the significance of birth signs or of particular aspects of a person's handwriting. But this certainly does not mean that these techniques necessarily predict either personality or the future. This is assessed by the psychometric principle of validity. In order to assess validity we first need to be clear about the purpose of an assessment. There are various forms that validity can take, the primary ones being face validity, content validity, criterion-based validity and construct validity. Face validity is the extent to which an examination or test 'feels right' for the person taking it. In a personality test, for example, are the questions actually relevant to the stated purpose of the test? Or in an examination does the type of question reflect the social world of the candidates, or is it alien to them? Content validity can be demonstrated by matching the assessment specification to the curriculum that has been followed. If candidates are set examination questions that are outside the syllabus, then this represents a failure in content validity. Criterion related validity is important when a test or examination is used to predict future performance on some criterion. The validity of school examinations for university entrance, for example, can be assessed by following successful candidates throughout their university life. Construct validity attempts to analyse what it is that a test or examination is actually measuring. It requires many years of research in which underlying issues fundamental to a concept are addressed in many different ways and from many angles. Differences of approach concerning the curriculum, pedagogical method and assessment of key aspects of schooling such as learning the 'times tables' or the phonetic approach to reading illustrate the struggle to define the constructs of ability and achievement in mathematics and reading in a meaningful way.

The third psychometric principle we need to address is standardisation. There are, in essence, two forms of standardisation: norm-referencing and criterion referencing. In practice there is often a complex mix of the two in public systems, as Newton and Baird remind us. A norm-referenced test or examination compares the score of an individual with those of other candidates who took the test under similar circumstances. This group of candidates is called the norm group. The ultimate norm group would be the whole population of potential test takers. The standardisation of the WIAT (Wechsler Individual Achievement Test), for example, was carried out by obtaining a stratified random sample of 800 children throughout the UK. The proportion of children chosen mirrored those in the 2001 Census in terms of gender, ethnic group, locality and parent's educational level. The use of this type of referencing is important when educational psychologists need to match the achievement of an individual child against reasonable expectations based on how children of a similar age achieve in the population at large. Criterion referencing refers to a matching of a test score or an examination result to some objectively assessed reference point that details how a person with this score might be expected to perform in future training or in the workplace. Some have attempted to set these forms of standardisation against each other, but such arguments are usually vacuous as both are important in most real world settings, each in their own way.

The final psychometric principle is freedom from bias. Bias occurs when scores on a test vary depending on group membership. A test, examination or assessment procedure is said to be biased when its use results in adverse impact on one or more groups when compared with others. Groups can be defined in many ways, but becomes particularly significant in areas where anti-discrimination legislation is in force, such as gender, ethnicity, social circumstance, disability, sexual orientation and now, age. There are three principle types of bias: item bias, intrinsic test bias and extrinsic test bias. Item bias occurs when some items within a test show group differences that are disproportionate with the test as a whole. It might occur, for example, where a particular item contains English that is far too colloquial when addressed to candidates for whom English is not their first language. Item bias is, in principle, fairly easy to identify, but much more could be done to ensure that procedures are in place to keep it to a minimum. Intrinsic test bias occurs where a test or examination has differential reliability or validity for different groups, and much of the research on intrinsic test bias was associated with attempts to introduce positive discrimination policies, particularly in the US. But latterly there has been an increased recognition that, apart from item level bias, most of the bias found in assessment is extrinsic to the test or examination itself. More often, differences in test scores between groups come about as a result of the impact of real differences in society. Bias in, and the consequent adverse impact of, school examination results can to a large extent be accounted for by differences between localities in the quality of schooling, or of parental, peer and teacher expectation and support. These are themselves dependent on the impact of social policy on local demographics.

How do the psychometric principles relate to the evaluation and development of school examinations such as the A-level? Very much. First, we need to dispel a common myth that A-level results are judgements, not measurements, and hence escape the need to be judged by psychometrics. Judgements, as much as measurements, need to be reliable, valid, well standardised and free from bias. Hence the principles are unavoidable. Furthermore, the distinction is in many ways artificial. Psychometrics today is defined as the science of psychological assessment, not simply measurement, and this is important, particularly when an organisation has to decide on how an assessment is to be made. In recruitment, for example, it is not simply a question of whether to use psychometric tests, interviews, or other alternatives such as work sample exercises. Rather, it is a question of comparing the reliability, validity, standardisation procedure and extent of bias that lie in each, and deciding on the overall package. To do this common criteria are needed and these the psychometric principles supply.

Politics and misunderstandings abound in the school examinations domain, and application of the psychometric principles enables us to divide the fact from the rhetoric in the frequent debates that are now part of our lot. Given what we know about reliability, how has it come about that we experience demands that examination results must be completely reliable, something we know to be impossible? The misunderstanding arises because all tests and examinations inhabit a world of conflicting meanings and interpretations, and therefore need to be assessed in terms of their consequences as well as their psychometric characteristics. In education these include progression, educational treatment, admissions, setting and streaming. Outside education, tests not only assess, they also license, and once test results are used to license they cross a threshold that interfaces with the wider legal and social system of society at large. Hence the award, for example, of a driving licence, or of membership of the Royal College of Surgeons, or of a place at University, give legal entitlements that, while based on assessment, achieve a new status that are inevitably going to be the subject of controversy.

To validate a public examination, as with any other test, we need first of all to define its purpose. This is a basic requirement as otherwise we could not know whether its purpose was being met. It is a multifaceted issue as each examination serves a number of different purposes, hence a series of validations are required. Problems can arise if some of these purposes are in conflict. For example, we may find that increasing validity in formative assessment may decrease validity in summative assessment. Furthermore, the simple knowledge that the result is being used for one purpose (e.g. school league tables) may decrease its validity for another. But, this said, there is no reason why an assessment should not serve a number of different purposes, so long as we are clear what these are, and where our priorities lie.

Standardisation is about standards, and there is an ongoing debate over whether standards, for example in A-levels, are going up or down. To get a grip on this we need to consider what is meant by 'standards'. For example, teaching standards are not the same as the standard of achievement. It is perfectly possible for standards of teaching to go up at the same time as standards of achievement go down, and vice versa. Also, standards are not necessarily applicable across the board. A form of teaching that raises standards for one group (for example, children with special educational needs) may lower them for another.

The desire to design assessments, examinations and tests that are free from bias is as much a concern for school examining bodies as it is for recruitment professionals. Unfortunately, given the existence of extrinsic test bias, assessment that is completely free from bias is in many cases an impossibility. But we can all endeavour to keep bias to a minimum, and to do so is an important part of any equal opportunities policy, whether that of an organisation or enshrined in law within equal opportunities legislation. What is important is that its extent should be monitored and discussed, and that programmes to evaluate and reduce its extent should be incorporated in policy. This can be difficult where companies and organisations are in denial, and it will be an uphill task to ensure that the issue receives the attention it deserves. As far as A-levels are concerned, two forms of bias are apparent. First, the differences in attainment between ethnic groups, and secondly, the superior performance of girls compared with boys, in some subjects. As far as ethnic groups are concerned, the differences in quality of schooling between inner cities and the suburbs is sufficiently manifest not to need much discussion, although the causes of these differences are of course a different matter. One thing we can be sure of, however, is that attempts to deflect the issue on to universities are unlikely to lead to the changes we need. The black and Bangladeshi communities in particular deserve to have their concerns in this respect recognised and addressed.

With gender differences in achievement, it is interesting to note that several decades ago boys outperformed girls at A-level, a situation that is now reversed. Is this because girls are now cleverer than boys? Not necessarily. Two other elements will almost certainly have come into play. First is the higher standard deviation for boys compared with girls on most ability and achievement tests. This generally means that boys are over-represented at the extremes of the distribution. A shift in the cut-off closer to the population average, as effectively happens when the participation rate shifts from 10% to 50%, could very easily show that the previous superior performance of boys was an artefact. A second change in the way A-level is examined will also have contributed, this being the increased dependence of the final mark on coursework. There are complex interactions between gender and various aspects of the coursework process.

The psychometric principles are not new, and necessarily underlie much of the activities of examination boards in their efforts to improve the culture of learning, examinations and the monitoring of performance. They are also inescapable, although sometimes attempts are made to dress them up in other clothes. Perhaps this is inevitable given the increasing politicisation of our school system. Is it too much to hope that one day the curriculum and its assessment will be disestablished? The freedom given to the Bank of England to set interest rates independent of Treasury interference has set a useful precedent here. Only time will tell.

VOCATIONAL ASSESSMENT

Is passing just enough? Some issues to consider in grading competence-based assessments

Martin Johnson Research Division

Introduction

Competence-based assessment involves judgements about whether candidates are competent or not. For a variety of historical reasons, competency-based assessment has had an ambivalent relationship with grading (i.e. identifying different levels of competence), although it is accepted by some that 'grading is a reality' (Thomson, Saunders and Foyster, 2001, p.4). The question of grading in competence-based qualifications is particularly important in the light of recent national and international moves towards developing unified frameworks for linking qualifications. This article is based on Johnson (2006, in submission) which uses validity as a basis for discussing some of the issues that surround the grading of competence-based assessments. The article is structured around 10 points taken from the summary of that extended paper.

1. Defining competency

This can be problematic and might be conceptualised in terms of atomistic/holistic or tacit/instrumental factors. Competency-based assessment systems have developed in the context of these varying conceptualisations.

The assessment systems used to represent and measure competent performance are inextricably tied to the ways that 'competence' has been defined. Debates about the nature of competence have tended to be polarised around the question of whether it is a complex or superficial construct, with consequent implications for assessment methods. Wood (1991) cites literature highlighting the inherent difficulties of inferring competence from test data or observed performance. He suggests that this is partly because those constructs that might be regarded by some as contributing to a notion of competency are often grossly under-conceptualised. This potentially leads assessment-based inferences about competence to be invalidly 'over-extended'.

More sophisticated conceptualisations tend to consider those attributes that underpin performance. Gonczi (1994) outlines a broad model of competence that prioritises the personally held skills which, in common, underpin competent performance. Gillis and Bateman (1999) also acknowledge a broader conception of competency, arguing that competency must include the application of skills across contexts (location and time), and the generic transferable skills, sometimes referred to as 'key skills', that enhance the capacity of workers to respond, learn and adapt when environmental factors change.

There are also concerns about whether competence can be satisfactorily defined and the role of assessor experience in judgements about competent performance. Some argue that attempts to over-specify detailed assessment criteria in order to attain unambiguous, reliable judgements might not have the desired outcome. Wolf (1995) observes that written specifications on their own might well leave space for ambiguous interpretation since no criterion, however precisely defined, is beyond multiple interpretations. Although it appears counterintuitive to suggest that very detailed assessment criteria may leave space for personal interpretation, when faced with a mass of criteria an assessor may well read through them and glean a sense of meaning, perhaps giving their own weight to particular points and therefore reducing the overall consistency of application.

Others argue that attempts to over-specify 'transparent' assessment criteria will also have limited success because of the particular influence of tacit knowledge in competent performance. Situated cognition theorists suggest that the development of competence involves 'knowing in practice' and becomes embodied in the identity of the practitioner (Lave and Wenger, 1991). Modelling the different stages of developing expertise, Dreyfus and Dreyfus (1986) argue that tacit, intuitive understanding is a critical difference between the performances of experts and novices.

2. Grading and motivation

There is considerable debate about the potential advantages and disadvantages of grading on motivation. Literature suggests that the reporting of performance outcomes can influence learner motivation. Social Cognitive theorists, such as Bandura (1986), hold that individuals use feedback from past experiences (successes and failures) to inform their expectations about future performance. Perhaps unsurprisingly, the quality of this information can affect perceptions of self-efficacy and influence future motivation to act.

Grading potentially gives more feedback about performance than binary reports. As a consequence, Smith (2000) suggests that grading can facilitate the motivation for students to strive for excellence since the reporting mechanism affords the opportunity for this level of performance to be recognised.

There is evidence that the effects of grading are not consistent across all learners. Williams and Bateman (2003) suggest that whilst more able learners might consider grading to be more motivational because it recognises their strengths, lower ability learners might be adversely affected. It is also important to consider the potential relationship between grading and labelling. There are concerns that learners might internalise the descriptive quality attached to grades to the extent that they infer that their performance (and ability) is a fixed, unchangeable entity.

The nature of learners who take vocational courses might be different from those who opt for general qualifications, and their motivation might differ. Group dynamic issues might also need consideration. Usually vocational learning takes place in smaller groups than is the case for general learning. This might contribute to a greater sense of group cohesion, undermining the motivation of individuals to compete against their peers.

3. The effects of grading on (mis)classification

Smith (2000) asserts that grading can improve the validity and consistency of assessments because it compels assessors to analyse students' performances with greater care than in binary reporting systems. This might be because they have to consider the evidence of a performance at a finer grain. On the other hand, this will only be possible if the inherent logic of the subject provides recognisable thresholds (Wolf, 1993).

Williams and Bateman (2003) highlight the potential relationship between the number of grading boundaries and the reliability of assessment outcomes. The opportunity for classification errors increases simply because the number of differentiated classifications increases. However, the errors might have less severe consequences. Overcoming this problem could potentially undermine the consistent reporting of outcomes since it demands greater levels of accuracy in each assessment judgement. Newton (2005) argues that the existence of measurement inaccuracy impacts on the social credibility of assessments because of public expectation that there should be relatively few misclassifications.

Finally, Wiliam (2000) emphasises the danger of aggregating marks into grades or levels since these might mask the true extent of error variance in test scores. Since the exactness of test scores can give an illusion of precision, resulting in misleading perceptions about their real accuracy, grading might be considered more favourable because it suffers less from this degree of definition.

4. Stretching assessment criteria beyond binary outcomes

Another important consideration is the interaction between domain breadth and the constructs included. Disentangling these interacting factors allows a clearer discussion regarding the potential consequences of grading. Domains can often be broad, requiring the integration of a number of identifiable skills. This raises questions about the nature of competent performance, since the term might be used (and understood) in different senses. Hyland (1994) suggests that competence might be both a holistic evaluation against a professional standard (e.g. being a competent plumber) and an atomistic evaluation of the ability to achieve a particular task (e.g. a particular driving manoeuvre). In the first context he argues that grading is appropriate because the holistic nature of the performance might include observable degrees of performance. However, in the second context grading might be inappropriate because atomistic tasks might not be scalable beyond 'achieved' and 'not yet achieved'.

5. Grading and accountability

There are concerns that grading procedures afford comparisons to be made between institutions and that these can be used for accountability purposes. In this way grading can be a potential source of pressure for assessors and might influence their decision-making. Wikström (2005) explores some of the structural pressures beyond the immediate context of assessment tasks which can impact on the integrity of grading decisions in a criterion-referenced system. She found evidence that teachers' and tutors' grading decisions were affected by selection and accountability concerns. Her findings suggest that teachers might grade differently over time because of:

> *Both internal and external pressures for high grading, due to the grades' function as a quality indicator for schools as well as a selection instrument for students*. (p.126)

Similarly, Bonnesrønning (1999) posits a systematic relationship between teacher characteristics, such as self-confidence levels, and grading practices. For example, he states that:

> *Teachers' ability to withstand pressure [for high grading] varies with teacher characteristics*. (p.103)

This could have implications for perceptions about the robustness of teacher or tutor assessed competency reports.

6. Decisions about grading depend on the domain being assessed

Decisions about grading need to consider the context of the domain being assessed. Grading decisions should be based on the number of usefully distinct subject specific criteria which can be formulated, the inherent logic of the subject, and whether there are recognisable thresholds. Messick (1989) argues that consideration of the consequences of assessment results is central to validity, stating that:

> *Validity is an integrated evaluative judgement 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*. (p.13)

Considering the interpretation of assessment evidence leads to a focus, amongst other things, on the quality of the information gained from an assessment. Although considering whether it is validly possible to separate binary into graded outcomes is important, arguably domains traditionally used for binary judgements grading can offer additional information. This could afford more information on which to base inferences about individual achievement and contribute to the validity of the assessment process. It assumes that inferences about competence are based on a sound understanding of the grading criteria. Transparency about how grades are determined is important. However, the meaning of different grade thresholds is less transparent than that between competent/not competent if there is a lack of understanding about the differences between grades.

7. Context and norm-referenced interpretations

Literature suggests that context and norm-referenced interpretations might undermine the validity of applying grading procedures to competency-based assessments. Context might interfere with consistency in at least two ways. First, context might interfere with an assessor's ability to position the qualities of two different performances on a common scale. Factors may exist that intrude on the process of casting consistent judgements (e.g. performances in tasks involving interactions between individuals might be interpreted differently by judges who accommodate variations in the social dynamics, such as, dealing with 'tricky' as opposed to 'helpful' customers). Secondly, context can make it more difficult to infer the basis on which assessors' decisions are being made. Assessors in different contexts might make judgements based on different foundations from each other because their understanding of competence is based on their different experiences.

Where binary reporting methods are used there is a clear, transparent link relating pass/fail distinctions to particular criteria. One of the problems for competency-based assessment is that qualification users might mistakenly assume that graded performance reporting is based on norm-referenced principles. Williams and Bateman (2003) and Peddie (1997) found that qualifications stakeholders sometimes make this mistake.

However, a number of commentators questioned whether criterion-referenced judgements are entirely devoid of norm-referenced principles. Skidmore (2003) argues that criteria could be based on an underlying normative judgement where they rely on subjective interpretation by professional judges. Similarly, Wiliam (1998), citing Angoff (1974), suggests that any criterion-referenced assessment is underpinned by a set of norm-referenced assumptions because the assessments are used in social settings, and assessment results are only relevant with a reference to a particular population. Consequently, any criterion-referenced assessment is attached to a set of norm-referenced assumptions.

8. The use of 'merit' grades

Using grading in competency-based assessments might demotivate and discourage some learners. One method of overcoming this problem is to grade outcomes once competence has been established. 'Merit' and 'excellence' grades might be used for this purpose, although Peddie (1997) suggests that these terms need to be distinguished so that they are used validly. According to Peddie, 'merit' and 'excellence' have different qualities; 'excellence' has an exclusivity, implying that some students are excellent in relation to a larger group of students who are not excellent, whilst 'merit' means very good, potentially being attained by all students. In this context, 'merit' grading can help to identify praiseworthy performances, without necessarily engaging the norm-referenced techniques that some argue undermine competency-based assessment principles.

9. Grading potentially affords the use of assessment data for selection purposes

An important use of assessment outcomes is to inform selection decisions, Wolf (1995) states a commonly held view that:

> *In a selection system, a simple pass/fail boundary provides far too little information on which to base decisions*. (p.75)

Grading can perform an important role where decisions need to be made about selection or access to limited opportunities and/or resources. Fewer grades will result in fewer fine distinctions between performance descriptions. The social consequences of this might be selectors placing a greater emphasis on other selection criteria, which might be less reliable than the examination/assessment itself. In addition, it reduces the effect of measurement error. For example, a pass might be a misclassified, incompetent applicant but an excellent result is much less likely to be one.

10. Grading can help to establish the comparative status of different qualifications

Grade creation based on the distribution of performances within the population can be one way of enabling comparisons between assessments to be made. It is also important to acknowledge that grading might encourage the use of particular frames of understanding which look to make comparisons across domains. The creation of graded performance scales might encourage the development of common assumptions about the similarity of skills and demands that are needed to achieve similar grades across different domains. The extent to which this is possible and valid is questionable although the construction of such comparisons is notionally encouraged by the grading framework.

A consequence of using grades as a tool for comparing the vocational and academic domains is the potential for 'a paradox of parity' (Griffin and Gillis, 2001). An important function of Vocational Education and Training (VET) is to encourage less academic students to remain at school. However, in order to achieve parity of esteem and intellectual demand with other 'academic' subjects, there is a perceived need to attract more academically able students into those vocational subjects. A paradox of parity could occur if this is successful since less able students might be discouraged from enrolling in VET courses which appear increasingly similar to 'academic' courses.

Conclusion

In theory, grading can be an appropriate method for dealing with ordinal competency assessment data, although there are claims that data from competency-based assessments should be regarded as being nominal. In practice, the potential benefits of grading need to be balanced against its potential disadvantages. This article suggests that questions about the desirability of grading competency-based assessments are related to issues of validity, with the question hinging on the simultaneous existence of two mutually supporting factors: 'use value' and 'validity'. It appears that the grading of competence-based assessments can only be justified where both factors exist, in other words where it has a clear value for qualification users and where its application is valid. The existence of either of these factors in isolation undermines the use of grading since it weakens the crucial link between the generation of sound assessment data and its complementary interpretation.

References

Angoff, W. H. (1974). Criterion-referencing, norm-referencing and the SAT. *College Board Review*, **92**, 2–5.

Bandura, A. (1986). *Social foundations of thought and action: A Social Cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bonnesrønning, H. (1999). The variation in teachers' grading practices: Causes and consequences. *Economics of Education Review*, **18**, 1, 89–105.

Dreyfus, H. L. & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the age of the computer*. Oxford: Basil Blackwell.

Gillis, S. & Bateman, A. (1999). *Assessing in VET: Issues of reliability and validity*. Kensington Park, South Australia: NCVER.

Gonczi, A. (1994). Competency-based assessment in the professions in Australia. *Assessment in Education*, **1**, 1 27–44.

Griffin, P. & Gillis, S. (2001). *Competence and quality: Can we assess both?* Paper presented at the National Conference on Grading and Competency Based Assessment in VET, Melbourne, May, 2001.

Hyland, T. (1994). *Competence, education and NVQs: Dissenting perspectives*. London: Cassell Education.

Johnson, M, (2006). Grading in competence-based qualifications: Is it desirable? *The Journal of Further and Higher Education*. In submission.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement*, 13–103. Washington DC: American Council on Education/Macmillan.

Newton, P. E. (2005). The public understanding of measurement inaccuracy. *British Journal of Educational Research*, **31**, 4, 419–442.

Peddie, R. A. (1997). Some issues in using competency based assessments in selection decisions. *Queensland Journal of Educational Research*, **13**, 3, 16–45. http://education.curtin.edu.au/iier/qjer/qjer13/peddie.html

Skidmore, P. (2003). *Beyond measure: Why educational assessment is failing the test*. London: DEMOS.

Smith, L. R. (2000). *Issues impacting on the quality of assessment in vocational education and training in Queensland*. Brisbane: Department of Employment, Training and Industrial Relations.

Thomson, P., Saunders, J. & Foyster, J. (2001). *Improving the validity of competency-based assessment*. Kensington Park, SA: NCVER.

Wikström, C. (2005). Grade stability in a criterion-referenced grading system: the Swedish example. *Assessment in Education*, **12**, 2, 125–144.

Wiliam, D. (2000). Reliability, validity, and all that jazz. *Education 3–13*, **29**, 3, 9–13.

Wiliam, D. (1998). *Construct-referenced assessment of authentic tasks: Alternatives to norms and criteria*. Paper presented at the 24th Annual Conference of the International Association for Educational Assessment, Barbados.

Williams, M. & Bateman, A. (2003). *Graded assessment in vocational education and training*. Kensington Park, South Australia: NCVER.

Wolf, A. (1993). *Assessment issues and problems in a criterion-based system*. London: Further Education Unit, University of London.

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

Wood, R. (1991). *Assessment and testing: A survey of research*. Cambridge: University of Cambridge Local Examinations Syndicate.

The Psychometrics Centre at Cambridge Assessment

Professor John Rust

The Psychometrics Centre is a centre of excellence within Cambridge Assessment, dedicated to the furtherance of scientific rigor in both pure and applied aspects of psychological measurement and educational assessment. We believe our position within Cambridge Assessment will help us to create links between educational and business assessment, leading to better practice, greater efficiency and social benefits.

Before moving to Cambridge, and after working at a number of academic institutions in London, Professor John Rust set up the Psychometrics Centre at City University, London, in 2003. This was the UK's first University centre dedicated to the field and John Rust became the UK's only Professor of Psychometrics. At that time, psychometric testing, while increasingly important in education, industry and the health sector throughout the world, was experiencing major new challenges as an academic and applied discipline.

The major challenge continues to be the internet, which has revolutionised the area, particularly in the field of recruitment testing. While as recently as 2004 it was possible to collect data for test development trials and norming studies by paper and pencil, customers in the business world now see such a suggestion as rather recherché. The migration of tests to the internet is accelerating fast, and demands from national professional societies as well as the International Test Commission that full revalidation is required before transferring from pencil and paper to computer, let alone the internet, has fallen on deaf ears. While only a few years ago it was argued that older candidates would be disadvantaged once tests were computerised, now it seems there is a reverse concern that younger candidates will be disadvantaged if they have to put pen to paper! The internet has also dissolved national barriers, and this is having a huge impact on the test publication and licensing process. Most published psychometric tests, particularly those used for diagnosis and recruitment, have traditionally only been sold to appropriately qualified professionals. But once on the internet, tests can be accessed from anywhere, and on a worldwide scale. Who is to decide who is appropriately qualified? The chances of international agreement on this are just as unlikely as those for the worldwide international school curriculum.

In the field of educational assessment, many of these issues are now just beginning to loom on the horizon. The US is in the lead, followed by Europe, but there is time to plan as changes will lag behind the introduction of broadband, which is a minimum requirement for serious internet delivery. The debate is but beginning, and there will be many who argue, often for sound reasons, that internet testing of children should not be permitted. But our experience in the recruitment testing world suggests that it is impossible to hold back the clock. We cannot ignore the fact that migration of assessment to the internet introduces a completely new business model that offers massive economies of scale – something our competitors will be well aware of. But while competition cannot be ignored, there is more to this than economics. While some cherished practices and values may fall by the wayside, the internet offers many new opportunities for the improvement of assessment that once were beyond our wildest dreams.

While valuing its contribution to testing in the business world, particularly in the field of human resources, most of the Psychometrics Centre's work continues to fall within the educational arena. Over the past 10 years John Rust and the Centre have been the developer commissioned by Harcourt Assessment (formerly The Psychological Corporation) to carry out the Anglicisation and UK standardisation of some of the world's most widely used diagnostic tests. These include the Wechsler tests, such as the WISC (Wechsler Intelligence Scale for Children) and the WIAT (Wechsler Individual Achievement Test) both widely used by educational psychologists for the diagnosis of special education needs, and the CELF (Clinical Evaluation of Language Fundamentals), used in a similar way by speech and language therapists. These projects have led us to collaborate with a wide selection of schools, preschools and nurseries throughout the UK, as well as over 100 professional Educational Psychologists and Speech and Language Therapists, who have been tasked with going into these schools and administering over 5,000 tests individually to children within a stratified random sample based on the Census data. We are currently embarking on such a project for the UK standardisation of the Ravens Progressive Matrices, a non-verbal test of ability.

The Psychometrics Centre has brought to Cambridge many of its activities previously carried out in London. For those who knew us there it is still 'business as usual'. The Centre will continue to act as developer and adaptor of tests; to deliver training, including the British Psychological Society's Level A and B certificates in test use that improve practical skills and technical understanding for HR professionals; to undertake applied research for the evaluation of assessment programmes, and pure research into key topics relating to web-delivered assessments, test adaptation across cultures and languages and predictive statistics; to offer specialist test-based consultancy to commercial and not-for-profit organisations; and to facilitate communication between different areas of assessment.

Unfortunately the MSc in Psychometrics that we ran previously at City University has had to be discontinued; however we are now working to create a much-needed programme of postgraduate study in the discipline within the University of Cambridge. The first part of this will be delivered within the MPhil programme of the Faculty of Social and Political Sciences this Michaelmas. We have also recruited our first Cambridge based PhD student who joined us in October, 2006. The Centre's move to Cambridge in November 2005 offers huge opportunities to develop our work through the international links, expertise and research capabilities of Cambridge Assessment and the University of Cambridge.

Research News

Statistical reports

Examinations generate large volumes of statistical data (approximately 800,000 candidates sit general qualifications each year). The objective of this series of reports is to provide statistical information about the system.

It is intended that the reports will focus on different aspects of the examination system and will be produced at a rate of two or three a year. In the first few years the issues addressed will vary from year to year but it is intended that some issues will be revisited, particularly if there has been a relevant change in the system.

Although this is a new report series, statistics of the examination system have been reported in various journal and conference papers listed elsewhere on the Cambridge Assessment website.

- Statistics Report Series No. 1: 'Provision of GCE A-level subjects'
- Statistics Report Series No. 2: 'Provision of GCSE subjects'
- Statistics Report Series No. 3: 'Uptake of GCE A-level subjects in England, 2001-2005'

http://www.cambridgeassessment.org.uk/research/statisticalreports/

Conferences and seminars

New Statesman

In July Sylvia Green participated in a roundtable discussion on 'Smart learning for the future' organised by the *New Statesman*. The session was chaired by Barry Sheerman, Chair of the Education and Skills Select Committee and speakers included Andrew Adonis, Minister for Schools. Participants discussed how far technology is a part of smart learning and how far we have gone towards smart assessment and recognition of the skills that it allows us to develop. The discussion was subsequently reported in the July 24th issue of the *New Statesman*.

The Society for Multivariate Analysis in Behavioral Sciences

John Bell attended the 25th Biennial Conference in Budapest in July and presented a paper on 'Modelling the predictive validity of selection tests'.

United Kingdom Literacy Association

In July Gill Elliott and Nat Johnson were invited to present their work on 'Variations in Aspects of Writing in 16+ examinations between 1980 and 2004' at a special session at the UKLA annual conference. The theme of the conference was *Teaching Reading and Phonics: Implications of the Rose Report*.

International Test Commission

John Rust and Vikas Dhawan attended the ITC 5th International Conference on Psychological and Educational Test Adaptation across Language and Cultures in Brussels in July. John Rust presented a paper on 'A multi-method approach to cross cultural test adaptation: a focus on qualitative methods'.

British Educational Research Association, University of Warwick

In September colleagues from the Research Division presented 10 papers at the British Educational Research Association Annual Conference. The papers reflected the wide range of research undertaken by the Division and will be discussed in a later issue of *Research Matters*.

Cambridge Assessment Conference

The second Cambridge Assessment Network Conference took place at Robinson College, Cambridge, on 16th October. The conference addressed the important issues surrounding how to assess students' abilities, following Professor Robert Sternberg's proposal that assessment should not only focus on an individual's ability to analyse, but also on their creative and practical skills. The main speakers were Professor Robert Sternberg, Dean of the School of Arts and Sciences and Professor of Psychology at Tufts University, Dr Ruth Deakin-Crick, Senior Research Fellow at The Graduate School of Education, University of Bristol, and Professor Peter McCrorie, Head of the Centre for Medical and Healthcare Education, St George's, University of London.

Association for Educational Assessment-Europe

In November Sylvia Green and Andrew Watts attended the 7th Annual AEA-Europe conference in Naples. The theme of the conference was 'Assessment and Equity'.

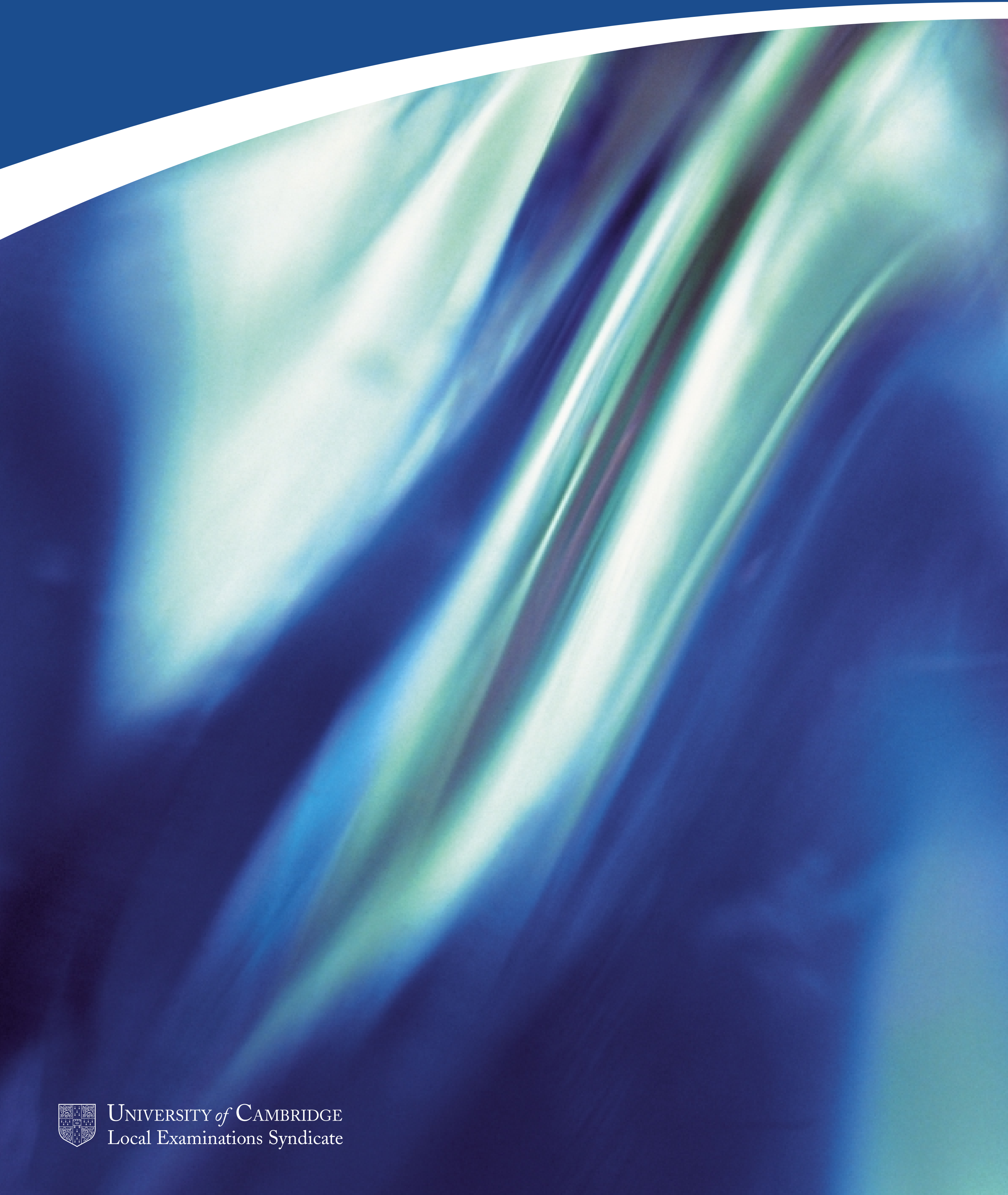
Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

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

Issue 3 January 2007

Research Matters

CAMBRIDGE ASSESSMENT



UNIVERSITY of CAMBRIDGE
Local Examinations Syndicate

CAMBRIDGE ASSESSMENT

Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate, a department of the University of Cambridge. Cambridge Assessment is a not-for-profit organisation.

Citation

Articles in this publication should be cited as: Crisp, V. and Johnson, M. (2006). Examiners' annotations: Practice and purpose. *Research Matters: A Cambridge Assessment Publication*, 2, 10-12.