Gold standards and silver bullets: assessing high attainment

John Bell Principal Research Officer, Evaluation & Validation Unit

One of the challenges facing those involved in the assessment and selection of high attainers is the fact that so many students get the same high grades (in measurement theory this is referred to as a lack of discrimination). For students who are concerned about their future opportunities the assessment of high attainment and the lack of discrimination at the top end of the ability range can be crucial. For example, the proportion of successful applicants to Cambridge who gained three grade As at A-level (excluding General Studies) rose to 93% (Cambridge University, 2005). However, of the 8,026 applicants who met the 3As criterion, 5,325 were not accepted (note the preference at Cambridge is for depth of knowledge in a small number of relevant subjects rather than breadth of knowledge in more subjects, so taking more than three A-levels is not necessarily regarded as desirable). As Geoff Parks, Director of Admissions for the Cambridge Colleges, noted, "Cambridge would prefer applicants thinking of stretching themselves, having chosen a coherent set of A-levels, to do so by stretching themselves 'vertically' by taking one or two Advanced Extension Awards rather than 'horizontally' by taking a further A-level". (http://www.cam.ac.uk/admissions/undergraduate/info/statements/pallisreview.html)

Of course, not every A-level student with three grade As applies to Cambridge. In 2004 there were 21,101 eighteen-year-olds with at least three grade As at A-level (excluding General Studies). This number represents 3.5% of all eighteen-year-olds and 7.9% of the eighteen-year-olds with at least one A-level result. This is one of the problems. From the perspective of admissions to elite courses the number is high, but from the perspective of assessment provision as a whole it is small. For individual subjects the problem is even worse. The numbers and percentage entries for various A-levels are given in the Table below. When the A-level content is central to a higher education course, then there may be many more applicants with a grade A in the required subject than places.

A-level Subject	*Number with grade A*	*% with grade A*
Biology	11,511	24.8
Chemistry	11,289	31.8
Physics	8,217	29.6
Mathematics	20,093	40.0
Further Mathematics	3,433	60.1
Geography	8,346	24.7
History	10.723	25.0
Economics	5,476	31.7
English	12,846	25.9

Source: 2004 Inter-Awarding Body Statistics

A-levels were introduced in 1951 for the purpose of determining who should be admitted into a limited number of university places. The situation is different now. Government policy is to increase the numbers in higher education and one consequence is that a much larger number of students have been allocated to the same number of grades. This means that there are more applicants for higher education with identical levels of performance when classified by grade combinations.

Something that happens in every field of human endeavour, from being a mechanic to a rocket scientist, is that casual observers are all too willing to offer simple solutions to problems, forgetting H.L. Menken's metalaw, 'For every human problem, there is a neat, simple solution; and it is always wrong'. In the case of education, the simple solutions are often wrong because they often fail to solve the problem. To many people, the objective seems to be partitioning the highest performers so that there is a smaller, more manageable group from which to select. However, the problem is not one of partitioning but one of measurement. The criteria that need to be considered are:

- What is meant by high attainment and who has it?
- Does the assessment predict future performance?
- What are the implications for fair access?
- What impact will this have on learning in schools and colleges?

Research is needed to understand what exactly is meant by high attainment and the answer may differ for different subject areas. When deciding on a method for identifying high attainment it is necessary to consider what exactly is being rewarded. The usefulness of selection tests can be summed up by the title of Dorothy Field's song, 'It's not where you start, it's where you finish'. The objective of selection procedures is to identify those applicants most likely to succeed. Of course, it has to be recognised that a selection test can only measure some determinants of performance. Implications of the assessment for fair access need to be considered to ensure that the assessment does not create unnecessary barriers to admissions. Students have to have equality of opportunity in demonstrating their level of attainment and their potential to succeed in higher education. For high attaining candidates the stakes are high and this means that 'open' assessments, such as coursework and dissertations, might not be perceived to be fair. Finally, a high stakes test is bound to have an impact on what happens in schools and colleges so it is important to recognise the wide range of purposes of school examinations besides selection for higher education.

At high levels of attainment the issue is not usually one of standards, rather it is one of selecting the best students. It is not usually the case that the applicants will not be able to cope with the demands of the course but that some will achieve more. The selection process is not about meeting a standard but about being the most suitable candidate. At lower levels of attainment, standards are more of an issue and involve the idea of minimal competence. It is also useful to recognise that particular A-levels may have different uses in different selection circumstances. These can be interpreted as evidence of attainment and evidence of potential. In the former, the A-level content is relevant and is used as a foundation for further learning and in the latter case, it is used as evidence that the candidate can cope with learning in a new area. In practice, this dichotomy is an oversimplification and both are relevant but the weight attached to each varies. The consequence of this is that responses to changes to the examination system will vary depending on who is using the examination results and for what purpose.

One result of the increased numbers with the same combination of grades is that there is either a lack of compensation (when a good performance in one subject is allowed to make up for a poor performance in another subject) or a reliance on performance in subjects of low relevance. For example, a student takes History, English, and Music A-levels and applies to do History at University. The offer is History A, English A, Music B but the student gets History A, English A, Music C. The student cannot compensate for the performance in music by showing an exceptionally high level of attainment in History. Who is more likely to be the better student, one who would have got an A++ History grade if one existed and a grade C in music, or a borderline History grade A and a grade B in Music? This issue of compensation is the major limitation of the new UCAS tariff (http://www.ucas.com/candq/tariff/index.html). There is no ceiling on the number of points than can be acquired and so compensation does not operate in a useful way. This means that a candidate who obtained only three grade As at A-level and had no other qualifications would have the same points as a candidate with three grade Cs at A-level and three grade Cs at AS. The latter candidate has not demonstrated any high level skills associated with a grade A. The two candidates are not equivalent.

The use of tariffs is further limited because it fails to take into account the requirements of particular HE courses. Hence, tariffs should not be used for evaluating whether the admissions process is fair (in terms of equal opportunities). For example, when considering those candidates with three grade As at A-level, including A-levels required for an individual course (i.e. most of the pool of suitably qualified candidates for a Cambridge course), the percentage of candidates meeting the criteria and attending independent (public) schools varies from 36.8% for courses requiring mathematics and physics to 51.4% for those requiring a language A-level.

Lotteries

Using the criteria listed above it is possible to consider various options that have been proposed to select high attaining candidates. One solution that has been proposed is the use of lotteries. Nothing is being measured and they are based on the assumption that all students have an equal probability of success. This is simply not the case. Within grade A there is a considerable variation in performance on the assessments described below and this performance is related to success. In terms of fair access, differences between a statistical concept of fairness (having an equal probability of being selected) and the real word 'fairness' (may the best person win) is an issue. In America where lotteries have been used, objections have been raised when the weaker students in the same institution (so resourcing and background are not issues) are selected in preference to the stronger. Lotteries are not an acceptable solution.

Marks

Another popular common suggestion is to use marks. The first difficulty is that the raw marks that are awarded to candidates are not really useful. A-level examinations are made up of modules and the combinations of modules vary from candidate to candidate. This can be the result of option choices or of taking the modules in different examination sessions. With a few exceptions, it is very difficult to construct examination modules of equal difficulty from session to session without pre-testing. This means that the raw marks have to be converted to uniform marks so that a grade A is always worth 80% of the marks on any module and in any session. This Uniform Mark Scheme was designed to be an intermediate step in awarding grades. Unfortunately, the system is designed to be fair to candidates close to the grade A boundaries and has potential problems for higher levels of performance. There is also a problem with exactly what is being measured. To ensure that the rate of exchange between raw and uniform marks is the same just above and below the highest grade, a cap is introduced. In some circumstances, the cap is lower than the maximum raw marks and maximum UMS marks are awarded to all candidates with raw marks equal to or above the cap (all but extremely erratic candidates obtain a grade A even if one or more of their modules is capped). Unfortunately, this capping process is an issue at higher levels of attainment.

Additional grades above A-level

Obviously the above property of UMS marks also affects the introduction of additional grades above A-level. There is also the issue of what exactly is being measured. There are two approaches that can be adopted. The first involves setting new boundaries on existing papers and the second involves adding additional material. However, there are difficulties associated with both of these alternatives. In the first case, the problem is that the examinations are not necessarily designed to test higher level skills above grade A. In investigating the feasibility of introducing additional grades at A-level, it was found that in some subjects it was considered possible to do this with existing examinations but in others it was felt that additional material would be essential. In these circumstances, it was felt that it would reward conscientious and careful students but would not identify those with higher order skills. However, adding harder material to the examinations is not straightforward because it alters the measurement characteristics of the examination as a whole. This is manifested either in very low grade E boundaries, a compression of the mark range between the A and E boundaries, or a lengthening of the examination, increasing the assessment burden. It should also be noted that only a small percentage of the candidates would be likely to complete the task satisfactorily because of its increased difficulty.

It is possible to investigate some consequences of extra grades using existing A-levels. For example, introducing A+ and A++ grades at equally spaced intervals gives the following results. For one OCR mathematics specification, 74% of candidates attended state schools but only 63.1% of those obtaining a grade A attended state schools. For the hypothetical A+/A++ grade the percentages are 55.3% and 49.1% respectively. Similar patterns were found for a range of other A-level subjects with pupils from independent schools increasingly represented in the higher range of the mark distribution.

Module grades

Another option is to use the grades obtained on each module. To be fair this would require all A-levels to have the same module structure. Even if this were the case then the tendency is to reward consistency. For example, consider two candidates with the following module grades (in lower case) and UMS marks (in brackets):

Candidate X:	a(80), a(80), a(80), a(80), a(80), a(80)
Candidate Y:	a(100), a(100), a(100), a(100), a(100), b(79).

If only the module grades are known and used then candidate X seems better but the UMS marks indicate that Y is almost certainly the better candidate. Obviously this is an extreme example but deciding whether a consistent performance is better than an erratic performance is debatable as it depends on the circumstances. Consistency is important for airline pilots as being brilliant at take-offs but useless at landings does not make a satisfactory pilot, but the history of the arts abounds with individuals classed as great for some of their work that was brilliant, even though much of what they did was less outstanding.

Additional assessments

There are other alternatives involving additional assessments. These can be grouped into three types:

- Subject specific examinations, e.g. Special papers and Advanced Extension Awards;
- HE course specific, e.g. the BioMedical Admissions Test (BMAT);
- General tests of high order skills, e.g. thinking skills assessments.

These tests have their own particular advantages and disadvantages. All of them have the advantage that they can be targeted solely for high attainers but there are also disadvantages in that they lead to additional costs and an increased assessment burden on candidates.

Subject specific examinations have the advantage that they are based on a particular subject area so that the measured issues can be addressed without compromise. However, there are access issues. In 2004 the uptake of Advanced Extension Awards was relatively low, from just 2 candidates (Irish) to 1,501 candidates (English). Universities cannot use them for making admissions' offers unless they are available in all schools so that they can be made a requirement. It can also be argued that such examinations favour schools and colleges that are either large, highly selective or well resourced because such institutions can provide the most effective support for candidates entering such examinations.

Another option takes the form of tests designed for admission to specific courses. This has the advantage that only relevant skills and knowledge are assessed. This can be a subset of the content of an A-level specification and can also include skills not directly assessed at A-level. In addition, all applicants can take them whether they enter A-level examinations or not and so they provide a common reference point for making decisions. These tests provide a way of assessing the potential of students whose ability might not be reflected in their grades. The objections relate to the extent that performances can be improved by coaching. This is a difficult issue. For example, it is accepted that the skills used in the BMAT will improve with familiarity and practice. However, it can be argued that these skills are really worthwhile, useful in many walks of life, and very important for success in higher education, and in this case it is possible for anyone to practice the skills involved with the help of freely and publicly available materials which are listed on the BMAT website.

Finally, there are tests that measure skills that are not directly assessed or not assessed by general qualifications. These skills may be important to success in higher education but it is important that the predictive validity is established. There has been some UK research into this issue (MacDonald et al, 2001a, 2001b). In a trial of the American College Board's SAT it was found that a different subset of high attaining candidates would be identified by the SAT compared with A-level. However, the study could not address whether these candidates would perform better than those identified by A-levels. The issue of predictive validity was not addressed.

One problem with experimental research is that it is low stakes. The outcome of the test has no impact on the future of the candidates taking it. The schools of the students in the experiment were not making any effort to prepare candidates for the test. If an aptitude test such as the American SAT were used for admissions, then this situation would change. For example, the BBC correspondent, Mike Baker, reported that "At Lafayette High School in Williamsburg, Virginia, I sat in on an 'SAT preparation' class. It's a sign of how important the SAT is in a teenager's life that this runs for 90 minutes a day for a whole semester. Piled high in the corner were some of the many preparation text books available on the market." (http://news.bbc.co.uk/1/hi/education/3304459.stm). It is for this reason that it is important that any general admissions test developed for higher education admissions should be developed to assess aptitudes that are educationally important and have long term benefits. Extensive research shows that coaching on aptitude tests has a small effect (Powers and Rock, 1999). However, this research was carried out to counter claims of coaching made by commercial providers. It does not mean that educational experiences do not influence the SAT score. For example, one of the College Board's researchers, Howard T. Everson, with a colleague, Roger E. Millsap, from Arizona State University (2004), investigated influences on SAT performance. They found that family background, learning opportunities in and outside the school curriculum and school characteristics influenced the SAT score.

Conclusions

In this paper, a number of options for assessing a small but important subset of candidates have been considered. UCLES has wide experience of all of these options and UCLES' researchers have conducted and will conduct many research projects that investigate the effectiveness of them. In particular, new methods are being developed to investigate the crucial factor of predictive validity. This research will be described in a future issue.

Whatever method of assessment is used, its effectiveness depends on how well it predicts future behaviour. This varies with circumstances so there is no simple gold standard that always identifies the best candidates in all circumstances. Neither is there a silver bullet – a test that measures the candidate's potential uninfluenced by an individual candidate's education experiences and personal circumstances. It is unreasonable either to expect this or to claim it of any educational assessment. However, it is important that assessments designed for this purpose do not add any biases and that they identify the candidates most likely to succeed. In addition, any preparation for the test should have a beneficial effect on the candidate, equipping them with skills that they will need as they progress through life.

References

Boyle, C. (1998). 'Organisations selecting people: how the process could be made fairer by the appropriate use of lotteries (with discussion)', *Journal of the Royal Statistical Society: Series D: The Statistician*, **47**, 2, 291–322.

Everson, H.T., and Millsap, R.E. (2004). 'Beyond individual differences: exploring school effects on SAT scores', *Educational Psychologist*, **39**, 3, 157–172. With correction **39**, 4, 261–261.

McDonald, A.S., Newton, P.E., Whetton, C. and Benefield, P. (2001). *Aptitude Testing for University Entrance: A literature review*. Slough: NFER.

McDonald, A.S., Newton, P.E. and Whetton, C. (2001). *A pilot of aptitude testing for University Entrance*. Slough: NFER.

Powers, D.E., and Rock, D.A. (1999). 'Effects of coaching on SAT I: Reasoning Test Scores'. *Journal of Educational Measurement*, **36**, 2, 93–118.

AUTOMATIC MARKING

Automatic marking of short, free text responses

Jana Z. Sukkarieh[1], **Stephen G. Pulman**[1] and **Nicholas Raikes**[2]

Introduction

Many of UCLES' academic examinations make extensive use of questions that require candidates to write one or two sentences. With increasing penetration of computers into schools and homes, a system that could partially or wholly automate valid marking of short, free text answers typed into a computer would be valuable, but would seem to presuppose a currently unattainable level of performance in automated natural language understanding. However, recent developments in the use of so-called 'shallow processing' techniques in computational linguistics have opened up the possibility of being able to automate the marking of free text without having to create systems that fully understand the answers. With this in mind, UCLES funded a three year study at Oxford University. Work began in summer 2002, and in this paper we introduce the project and the information extraction techniques used. A further paper in a forthcoming issue of *Research Matters* will contain the results of our evaluation of the automatic marks produced by the final system.

Uses for automatic marking

UCLES' traditional strength is in high stakes assessments that lead to qualifications. As more of our customers move to computer based assessments, an initial application of automatic free text marking in a high stakes context is as a quality control check on human marking, increasing the speed and efficiency of our quality control process. Every short, free text answer[3] could be marked both by computer and human markers, with any differences being resolved by a second human marker. Over time, as the capabilities and limitations of automatic marking became better understood, the proportion of answers marked by both computer and human could be reduced, with human marking targeted on the hardest to mark questions and on reviewing automatic marks that appear anomalous.

In the short term, however, the real opportunity for automatic free text marking is in low stakes tests. Many teachers and students use questions from our past papers, and we would like to be able to offer them an automatic marking service covering the free text questions as well as the 'objective' ones.

The challenge

Raikes and Harding (2003, p.270) state that an item's suitability for automatic marking depends on how near it can be placed to the objective end of what they call the objective-subjective continuum. The continuum is defined by the 'resolution' – the specificity and comprehensiveness – of an explicit marking guide that specifies how answers should be processed and marked. Traditionally, high resolution guides have been generated by greatly constraining the answers that students may give, as in multiple choice tests. More recently, attention has focussed on techniques for generating what are in effect high resolution marking guides for more open-ended item types, shifting them towards the objective end of the continuum where they may be automatically marked without affecting their validity.

In our automatic marking project we were concerned with marking short, factual answers varying in length from a few words up to around five lines, taken from GCSE biology examinations, where answers were marked for their correct content. The challenge was in coping with the myriad and sometimes unconventional ways in which credit-worthy answers were expressed, and the many mistakes in grammar and spelling found in some answers that nevertheless contained more or less the right content. Standard syntactic and semantic analysis methods would have been difficult to use, and even if we had fully accurate syntactic and semantic processing, many answers contained features that require a degree of inference that is beyond the state of the art. For example, in a question concerning asexual reproduction, a human marker inferred that

[1] Computational Linguistics Group, Centre for Linguistics and Philology, Walton Street, Oxford OX1 2HG, United Kingdom. Email: first.lastname@clg.ox.ac.uk

[2] Senior Research Officer, Evaluation & Validation Unit

[3] Of course, completely objective items – multiple choice items and the like where every possible answer can be predicted – will be marked solely by computer.