Issue 12 June 2011

Research Matters

CAMBRIDGE ASSESSMENT

University of Cambridge
Local Examinations Syndicate

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

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

Research Matters : 12

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

This *Research Matters* is published at a time of considerable change. Alison Wolf has completed her review of 14–19 vocational education and training. Dame Clare Tickell's review of the Early Years and Foundation Stage (EYFS) has reported. Reviews of the National Curriculum and National Assessment at KS2 are underway, the first overseen by an Expert Panel chaired by myself, and the latter chaired by Lord Bew. Inspection is being reviewed, and major changes in the governance of schools are being implemented. None of the articles in *Research Matters* appear to engage with these changes. But this is deceptive. The assessment community needs to remain focused on fundamental concerns relating to the purposes and uses, and technical characteristics of assessment – the research included here highlights continued attention to these 'bedrock' issues. The work on Diplomas – despite policy moves away from these – causes us to reflect on that key introductory phase of new qualifications – a perennial rather than a passing matter. But two articles link straight to the issue of change. The matter of what certification we require at the age of 16 is again a focus of debate – reflecting on how this issue was examined a century ago is salutary. The issue of how we compare and maintain qualifications in a changing and diverse system raises issues which are technical – how do we do it – and cultural – the maintenance of public confidence. In periods of change, it is vital not to lose sight of those things to which we constantly should attend.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

This issue takes us from matters related to examinations that were topical in 1911 through to today's processes and techniques and the development of new research methods. In the first article Elliott discusses a report on secondary school examinations from 1911 and discovers that many of the concerns expressed at that time resonate with current issues. This historical perspective from one hundred years ago is interesting and informative highlighting the fact that we can learn from the past and should do so. Shiell and Raikes move us to present day processes in the context of on screen marking and the analyses of item level data (ILD). The availability of ILD has enabled the routine production of reports on marking and grading and this article reports the results of a questionnaire survey that was conducted to evaluate how the reports were used by examiners. Crisp and Green then report on the evaluation of a new qualification and the practicalities of its implementation. Qualitative data were gathered in the first year of the 14–19 Diploma with research focusing on the Principal Learning part of the Diploma and providing insights from students and teachers about their perceptions across a range of themes. Although some problems were identified, the overall picture was positive and the findings highlight some important lessons to be learned for future developments of this kind.

Bramley and Dhawan's article is based on part of a report commissioned by Ofqual (the exams regulator in England). They investigated the challenging area of assessment reliability and explored what the impact of slightly different decisions at unit/component level would be on the grade distributions at whole assessment level. With the work of Shaw and Bailey we move to the complex area of predictive validity. In their article they emphasise how important it is to demonstrate predictive validity for tests that are used for university selection purposes. Establishing predictive validity poses a range of practical problems which must be overcome and the case study approach reported in this article serves to enhance understanding of future work in this field.

The first of two articles on research methods is by Johnson and Mehta. They review issues related to the comparison of demand and evaluate the CRAS framework (Complexity-Resources-Abstractness-and-Strategy). The authors discuss the use of the framework for individual items and the potential difficulty of using it at a holistic paper level. In the second article Greatorex, Mehta, Rushton, Hopkin and Shiell describe the process of developing a research instrument to compare diverse qualifications and subjects. This tool is particularly useful for comparability studies where it is difficult to use candidates' performance and assessment tasks, such as for new and vocational qualifications.

Sylvia Green *Director of Research*

Lessons from the past: An overview of the issues raised in the 1911 'Report of the Consultative Committee on Examinations in Secondary Schools'

Gill Elliott Research Division

> *We will improve standards for all pupils and close the attainment gap between the richest and poorest. We will enhance the prestige and quality of the teaching profession, and give heads and teachers tough new powers of discipline. We will restore rigour to the curriculum and exam system and give every parent access to a good school.*
> Conservative Party Manifesto, 2010

To what extent have the issues and concerns in education changed during the past century? The summary of educational aspirations provided in the Conservative Party Manifesto identifies seven key matters of direct concern to students, teachers and parents in 2011. However, how different are these topics to those which were investigated a century ago? This article summarises the educational issues raised by the 1911 report "Examinations in Secondary Schools" (Board of Education, 1911), a document which made recommendations which were to set in place an educational system in England which proved both enduring and successful, and examines briefly how many of the issues are still current today.

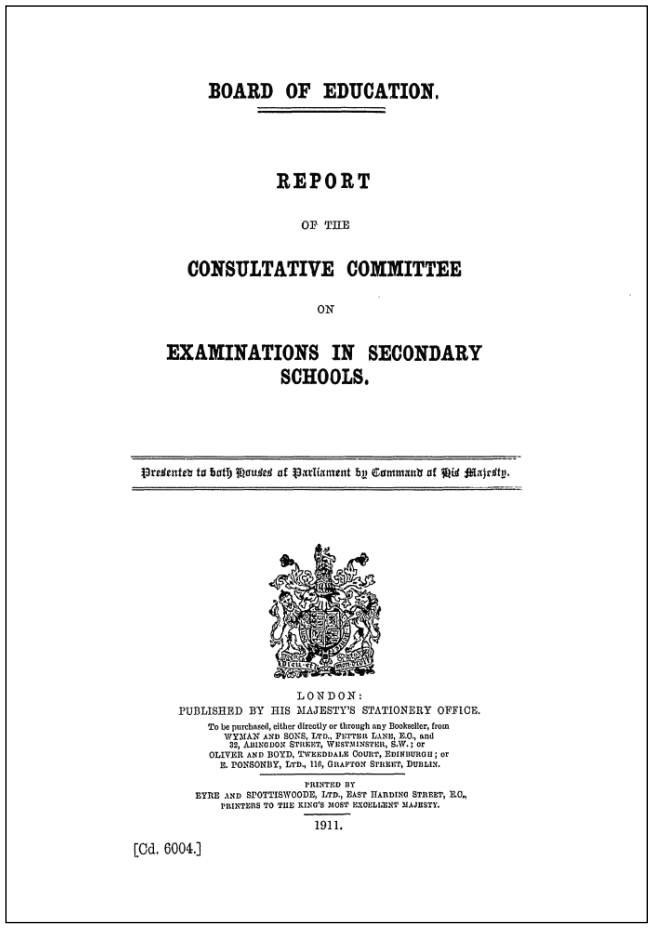
It is beyond the scope of the article to compare the issues with those of today in any great depth; rather the intention is to celebrate the centenary of the report with an overview, which it is hoped will allow other commentators to explore the material in greater detail. The temptation to 'pair' quotations from 1911 with those from more recent documents, either to illustrate the similarity of thinking or the diversity of approach, has been resisted as far as possible, despite the fact that the 1911 report contains so much detail that it would be possible to find examples of quotations for many current issues. Where documents are quoted, the title and date of the report are given as sources, rather than the authorship, as this avoids the jumble of acronyms and lengthy committee titles which would otherwise ensue.

"Examinations in Secondary Schools": The 1911 report

In 1911 the Board of Education invited a committee to consider the question of "when and in what circumstances examinations are desirable in Secondary Schools (a) for boys and (b) for girls."

The committee comprised twenty individuals, and was headed by the Right Honourable Arthur Herbert Dyke Acland, an Oxford-educated barrister who had been MP for Rotherham between 1885 and 1899 (UK Parliamentary Services, 2009). Amongst the other members of the committee were two church ministers, one professor, two doctors and an MP. A brief search into the background of some of the committee members shows that they comprise similar figures as might be invited to provide evidence to government today: amongst others were Marshall Jackman, who was Secretary of the National Association of Inspectors of Schools and Educational Organisers, Albert Mansbridge, who had founded the Workers' Educational Association in 1903, and Harry Reichel, who was instrumental in founding a national University of Wales (Aldrich and Gordon, 1989). The committee included four women. Three of these were Margaret Tuke, one of the first women to be educated at Cambridge, and Principal of Bedford College at the time of the report, Sophie Bryant, Headmistress of North London Collegiate School, suffragist, campaigner and mountaineer, and F. Hermia Durham, an historian, first winner of the Alexander Prize (Royal Historical Society, 1945) and between 1907 and 1915 the organiser of trade schools and technical classes for women for the London County Council (Hartley, 2003). In 1915 she was appointed to lead the programme of engaging women to keep businesses running during the First World War.

The political background to the 1911 report bore some similarities to that seen in 2010 and 2011. A General Election had been held in January 1910, after the House of Lords vetoed David Lloyd George's 1909



BOARD OF EDUCATION.

REPORT

OF THE

CONSULTATIVE COMMITTEE

ON

EXAMINATIONS IN SECONDARY SCHOOLS.

Presented to both Houses of Parliament by Command of His Majesty.

LONDON:
PUBLISHED BY HIS MAJESTY'S STATIONERY OFFICE.
To be purchased, either directly or through any Bookseller, from
WYMAN AND SONS, LTD., FETTER LANE, E.C., and
32, ABINGDON STREET, WESTMINSTER, S.W.; or
OLIVER AND BOYD, TWEEDDALE COURT, EDINBURGH; or
E. PONSONBY, LTD., 116, GRAFTON STREET, DUBLIN.

PRINTED BY
EYRE AND SPOTTISWOODE, LTD., EAST HARDING STREET, E.C.,
PRINTERS TO THE KING'S MOST EXCELLENT MAJESTY.

1911.

[Cd. 6004.]

'People's Budget'. The 'People's Budget' had sought to introduce new taxes on the wealthy (most notably a land tax and increased inheritance tax), the revenue from which was intended to bring about social reform through social welfare programmes. Instrumental in the budget were Liberals Asquith, Lloyd George and Churchill. The January election resulted in a Conservative/Liberal Unionist hung parliament. A second election was held in December, but produced an exact tie in results, and the Liberals formed a government with support from the Irish Nationalists.

Undoubtedly a time of great political and social change (especially with regard to the role of women in society, which is reflected in the particular detail given to girls' education in the various reports discussed in this article), a considerable number of Committees were commissioned in order to comment on matters of social concern. In the field of education alone there were six major investigations between 1906 and 1916, comprising 'Questions affecting higher elementary schools', 'School attendance of children below the age of five', 'Attendance, compulsory or otherwise, at continuation schools', 'Examinations in secondary schools', 'Practical work in secondary schools' and 'Scholarships for higher education'.

In the century since the publication of this report, many aspects of society have changed out of all recognition. Transport, is one example of this and telecommunications another. Edwardians, whilst present at the birth of the motoring and flight industries and well acquainted with railways, would undoubtedly be amazed by the extent, variety and speed of transportation infrastructure in place today. Equally, although the centennial anniversary of the first telephone was celebrated in 1976, the development of satellite systems, mobile telephones and internet has revolutionised the way in which we communicate. But what of education, and particularly, assessment? Would the Edwardian members of the 1911 Consultative Committee recognise the issues in assessment and testing which beset us today? Or have the changes in policy and practice which have occurred in the meantime altered the underlying concerns?

Between 1911 and 2010 at least 52 Acts of Parliament related to education were passed, informed by some fifteen White Papers. Admittedly some of them exist only to repeal the Acts of previous administrations; others still are minor, amending some small part of the system. Nevertheless, it seems likely that there has been substantial change in the system, given the amount of legislation that has been enacted.

The Edwardian drive towards commissioning investigative reports from individuals who might be expected to combine sound research skills with relevant expertise has remained a feature of the education system throughout the century, beginning with six Hadow reports between 1923 and 1933. Figure 1 provides a list of some of the reports which followed.

The titles of these reports give an indication of the vast breadth of interest that has been taken in education.

Returning to the 1911 Report into Examinations in Secondary Schools, to what extent are the specific concerns about assessment continuing to pose problems today? The report was organised into five chapters:

- A history of education in England
- A description of issues and problems, entitled "The Present State of Things"
- Further investigation of issues, entitled "The Difficulties and Disadvantages of the Existing System of External Examinations in Secondary Schools"
- Suggestions for reform
- Practical solutions.

Figure 1: Education Reports 1934–present

Spens (1938) *Secondary Education*
Norwood (1943) *Curriculum and Examinations in Secondary Schools*
Fleming (1944) *Independent Schools*
Percy (1945) *Technological Education*
Barlow (1946) *University Places*
Clarke (1947) *School and Life*
Clarke (1948) *Out of School*
Gurney-Dixon (1954) *Early Leaving*
Crowther (1959) *15–18 Provision*
Beloe (1960) *Secondary School Examinations other than GCE*
Newsom (1963) *Half our Future*
Robbins (1963) *Higher Education*
Lockwood (1964) *Schools Council*
Plowden (1967) *Children and their Primary Schools*
Newsom (1968) *Public Schools Commission*
Dainton (1968) *Science and Technology in Higher Education*
Donnison (1970) *Public Schools Commission*
Durham (1970) *Religious Education*
James (1972) *Teacher Training*
Russell (1973) *Adult Education*
Swann (1974) *The Flow into Employment of Scientists, Engineers and Technologists*
Bullock (1975) *A Language for Life*
Taylor (1977) *A New Partnership for Our Schools*
Waddell (1978) *School Examinations*
Warnock (1978) *Special Educational Needs*
Mansell (1979) *A Basis for Choice*
Rampton (1981) *West Indian Children in our Schools*
Cockcroft (1982) *Mathematics Counts*
Thompson (1982) *Youth Service*
Swann (1985) *Education for All*
Kingman (1988) *Teaching of English*
Higginson (1988) *A Levels*
Elton (1989) *Discipline in Schools*
Rumbold (1990) *Starting with Quality*
Dearing (1993) *The National Curriculum and its assessment*
Dearing (1996) *Higher Education in the Learning Society*
Kennedy (1997) *Further Education*
Moser (1999) *Improving Literacy and Numeracy*
Tomlinson (2004) *14–19 Curriculum and Qualifications Reform*
Steer (2005) *Learning Behaviour*
Steer (2009) *Learning Behaviour: Lessons Learned*
Rose (2009) *Independent Review of the Primary Curriculum*
Browne (2010) *Securing a Sustainable Future for Higher Education*

Table 1: The main issues identified in the 1911 report

Issue	*Details*
The role of the Board of Education and its relationship with awarding bodies	A lack of communication and co-operation between examining bodies and authorities, although the Oxford Delegacy and UCLES are praised for their Joint Board on behalf of Oxford and Cambridge Universities, as are the Universities of Liverpool, Manchester, Sheffield and Leeds. *... the Board of Education do not themselves conduct examinations in Secondary Schools (except indirectly, of course, by means of their Preliminary Examination for the Certificate), nor have they laid down any specific rules for external examination... Generally speaking it would appear that there is no formal co-operation between the Board of Education and the various examining bodies, so far as the actual conduct of their examinations is concerned.* The Committee reports that there has been some, limited, progress in terms of co-operation and knowledge-sharing between the Board of Education and the ABs through the work of the Board's schools inspectors, who are encouraged to comment upon the preparation for external examinations as witnessed in schools, and to share this with the ABs. Equally, the ABs are encouraged to supply copies of their reports on Secondary Schools to the Board. However, *in practice the actual extent of the co-operation is as yet somewhat slight. The examiners hardly ever inspect, and the inspectors never take part in external examinations, nor are their respective estimates of the general efficiency of a school ever officially correlated.*
Too many awarding bodies, all operating in overlapping areas	*The possibility of more concerted action under the conciliatory action and unifying influence of the Board of Education.* Leads to 'incidental' competition.
Equivalence of qualifications	*The diversity and independence of examining bodies make it impossible to find a common denominator between their examinations. The difficulty is not that the existing standards are too high or too low, but rather that those of different bodies vary, and that a recognised standard cannot at present be settled on its merits.* (Also III vii)
The use of examination results to enhance a school's reputation	*This point was put to us very plainly by Mr Cyril Norwood "Schools," he said, "were greatly tempted to produce as long an honours list as possible, and put boys and girls through examinations which were often quite unnecessary and even a hindrance. Sometimes clever pupils were utilised rather unscrupulously to enhance the credit of a school by achieving examination successes."*
Problems arising from the wide number of combinations of examinations and the way in which their comparability is used	*... we may point out that while candidates can obtain their Oxford Senior Certificate by passing in five subjects, no one set of five subjects is accepted by the exempting bodies. A candidate would have to pass in eleven subjects ...to be sure that his certificate would be accepted by all the bodies who accept the Oxford Senior Certificate as qualifying a candidate for exemption from their Matriculation or Preliminary Examination. If he only passed in the five subjects required by one particular body, and then for any reason changed his plans and needed to use his certificate to obtain exemption from the examination of some other body, he might find it quite useless to him...*
A multiplicity of examinations in schools	The Committee had made extensive efforts to gather data to evaluate the position. A survey from Lancashire suggested that approximately 26 different examinations were (commonly) taken. The data suggested that 1,070 students from this region entered examinations during the year 1910–1911 (just under a fifth of the 12–16 school population) and the ages of those students were as follows: 2 below 12, 38 aged 12, 112 aged 13, 169 aged 14, 261 aged 15, 314 aged 16, 230 aged 17, 106 aged 18 and 42 aged 19. Much information is presented about local regulations which were being brought in to forbid schools from presenting scholars for examination at the younger age ranges, and also regulating the number of general examinations which might be taken. For example, *The Middlesbrough Education Committee forbids pupils to take any external examination other than the Cambridge Local until they have entered their fourth year at school.*
Failure of many present external examinations to have regard to some important parts of school curriculum and school life	The examples given are vocational subjects *which cannot as a rule be tested without inspection, and that such inspection would be very costly even if the examining bodies had a staff of inspectors competent to do the work.* Moral and physical training, pupils' character, behaviour, steadiness, perseverance, influence, all omitted from external examinations.
School inspections	Described as a recent innovation. *A full assessment is held every 3–5 years, and an ordinary inspection every year.* The committee voices a concern that the inspection reports (a 'reasoned' report on the whole working of the school) are often not made available by the schools to the parents whereas examination successes are. The exam boards are accused of conducting both formal inspections, and using their position to carry out additional inspection: *sends examiners who, in fact, conduct what is a virtually an informal kind of inspection as part of their examining work. Thus...the work of their examiners includes visits to the school for the purpose of inspecting the buildings and apparatus, observing the school organisation and discipline, and hearing lessons given by school staff.*
The demands which examinations make upon the pupils' school time	*Mr Paton supplied us with definite instances in which pupils had spent nearly six weeks of their summer term in attending scholarship examinations...the loss of 30 per cent of their time which would otherwise have been given to systematic coherent study in class.*
Isolation of the examining bodies from the schools	This *causes problems with curriculum, school methods and school experiments* (e.g. subjects which are less easy to examine are left out of the curriculum, teaching methods are restricted to those which assist examination success and development of alternative types of school are hindered).

Table 1 summarises the main issues identified in the report, as described in the second and third chapters.

Other issues mentioned in less detail in the 1911 paper include:

- Premature disintegration of classes due to multiplicity of external exams.
- Teachers not having a large enough role in terms of consultation on external exams.
- No sound way in which schools may be judged by the public.
- Physical and mental overstrain [of pupils].
- Failure of the exam system to keep pace with educational innovation in schools.
- Special difficulty facing the Civil Service Commission because of the needs of international candidates from elsewhere in the Empire.
- Parental pressure.

- Awarding Bodies' class lists, honours, distinctions, prizes and scholarships accentuating the competitive element of examinations.
- Extent to which University requirements determine the syllabus of Secondary School examinations, though the number of pupils who proceed to University is a very small minority.

To what extent are these issues still current a century later?

The role of the Board of Education and its relationship with awarding bodies has changed greatly. Far from there being 'no co-operation and knowledge-sharing', there are strong links between the Regulator (Ofqual), the awarding bodies and other educational bodies. This was formalised in the Apprenticeship, Skills, Children and Learning Act (2009) where Ofqual's remit was defined:

> *It's our duty to ensure standards are maintained in the qualifications system. We primarily do this by evaluating qualifications, and the bodies that award them, against nationally established criteria. For this reason we formally recognise awarding organisations by checking they have adequate resources to award their qualifications.*

The argument that there were too many awarding bodies, all operating in overlapping areas has been partially addressed with the introduction of regulation, which was also a part of the recommendations of the 1911 Committee:

> *The establishment of a central Examinations Council, widely representative in character and entrusted with the powers necessary for carrying out the main principles laid down in this Report… The function of the council would be the supervision of all external examinations in recognised Secondary Schools… The establishment of an Examinations Council on such lines would secure in all essential points the advantages of centralised authority and of diversified experience, both of professional and local needs. It would bring into order the present confusion. It would replace multiplicity of standards by unity of control.* (Examinations in Secondary Schools, 1911)

Arguments about equivalence of qualifications still dominate educational forums, and the 1911 commentary on this is revealing – it was not that standards were too low or too high which was the problem, rather the committee identified difficulty in deciding upon an agreed standard. In some ways this has become more complicated in the present day, with difficulties deciding upon how to define a standard, let alone set in place its agreed 'merits'. This has, to a large extent been brought about by the expansion of purposes to which the results of examinations are put, and brings us to a situation which is very similar to that described in 1911: *problems arising from the wide number of combinations of examinations and the way in which their comparability is used*. In 1911, students wishing to follow different pathways into further training or employment needed to take multiple sets of examinations; in 2011 there are widespread questions about the suitability of the available assessments to sufficiently fulfil the different purposes to which they are put. Whilst the issue of *a multiplicity of examinations in schools* does not necessarily exist to the same extent in the context of age 14–19 public examinations (which was the 1911 context), it still exists in the arguments about National Testing, as described in Testing and Assessment (2008) as the 'burden of testing'.

Using examination results to enhance a school's reputation was a practice rooted in the behaviour of schools themselves, according to the 1911 report. In 2008:

> *…we find that the use of national test results for the purpose of school accountability has resulted in some schools emphasising the maximisation of test results at the expense of a more rounded education for their pupils. (*Testing and Assessment, 2008*)*

The concern in 1911 was that 'clever' pupils were subject to unnecessary examinations in order to reflect well upon the school. In 2011 the concern tends to be that schools direct more attention to the C/D borderline students and other students suffer.

> *…the focus of GCSEs has been very heavily on the C–D border line, and not, for example, on students underachieving by getting a grade A, but who could hopefully get an A*, or on those getting a B, but who could be helped to get an A.* (Testing and Assessment, 2008)

In both 1911 and 2008 there is concern about the narrowing of the curriculum, as can been seen by the similarity of the sentiments expressed in the two quotations below:

> *…there must always be a danger that young pupils will be allowed to drop useful but uncongenial subjects at too early an age, whether for their own supposed advantage or for that of the school.* (Examinations in Secondary Schools, 1911)

> *…the majority of time and resources is directed at those subjects which will be tested and other subjects in the broader curriculum, such as sport, art and music, are neglected.* (Testing and Assessment, 2008)

The 1911 concern about the *failure of many present external examinations to have regard to some important parts of school curriculum and school life* is perhaps the one issue least changed today, as the quotations below illustrate:

> *Tests, however, can only test a limited range of the skills and activities which are properly part of a rounded education, so that a focus on improving test results compromises teachers' creativity in the classroom and children's access to a balanced curriculum.*

> *The phenomenon described as 'narrowing of the curriculum' is strongly related to teaching to the test and many of the same arguments apply. There are essentially two elements to this concept. First, there is evidence that the overall curriculum is narrowed so that the majority of time and resources is directed at those subjects which will be tested and other subjects in the broader curriculum, such as sport, art and music are neglected. Second, within those subjects which are tested, the taught curriculum is narrowed to focus on those areas which are most likely to be tested ('narrow learning') and on the manner in which a component of the curriculum is likely to be tested ('shallow learning').* (Testing and Assessment, 2008)

The demands which examinations make upon the pupils' school time is apparent in recent arguments:

> *Another theme which manifests strongly in the evidence relates to the quantity of testing and there is concern that the quantity of national testing is displacing real learning and deep understanding of a subject.* (Testing and Assessment, 2008)

However DfES evidence to Testing and Assessment (2008) strongly opposed this, pointing to recent changes, including: KS1 testing

incorporated into normal lesson time, KS2 testing totalling less than 6 hours, KS3 testing totalling less than 8 hours, less coursework at GCSE and reduction of the number of A level units from 6 to 4. Additionally:

> *The Minister told us that no pupil spends more than 0.2% of their time taking tests*. (Testing and Assessment, 2008)

The issue of quantity of examination time also features in the linear-modular debate:

> *In addition, it is currently possible for AS students to sit retakes in order to maximise their grades at the end of the A-level course. It has been argued that this places too great a burden on pupils, diverting them from study of the course to focus on examinations*. (Testing and Assessment, 2008)

Isolation of the examining bodies from the schools was not described consistently in the 1911 report. On the one hand there was concern that the awarding bodies were not close enough to schools to be able to adequately provide, in the assessment curriculum, a true reflection of schools' needs. However, in the school inspections discussions, the awarding bodies were criticised as being somewhat over-eager. Watts (2008) confirms that inspection was considered a part of the examinations system and formal procedures existed for this. Fast-forward one hundred years and inspections (in England) are the remit of an independent, impartial non-ministerial government department. Similar departments exist in Scotland (HMIe), Northern Ireland (Education and Training Inspectorate) and Wales (Estyn). Whilst the role of school inspections has moved away from the awarding bodies, relationships with schools have strengthened greatly – support and training to schools from awarding bodies is available via formal events (such as Inset) and less formal means, including internet discussion boards and extensive support materials.

Some of the more minor concerns of the 1911 committee are still an issue today. Disruption of classes due to the multiplicity of external exams is no longer a problem in the sense understood in 1911, but does still feature in the linear-modular debate. The soundness of the means by which schools may be judged by the public remains a current concern, as does the balancing act of assessing a curriculum suitable for Higher Education needs whilst at the same time providing for students who do not intend to follow that route.

Looking at the seven key aims of the current government there is much that was of concern in the 1911 committee report, notably the issues of curriculum rigour, the system of examining and the improvement of standards. Whilst there are plenty of examples of instances where the issues have changed, even turned upside down, it is clear that were Arthur Dyke Acland and his fellow committee members to be presented with the issues at stake in 2011, there would be much that they would recognise from their deliberations in 1911. It is to be hoped that they would be pleased – much of the underlying structure of the current system, including development of the current GCSE and A level qualifications structure, has evolved from the antecedent qualifications structure suggested in their report. However, all three of the fundamental principles of the examination system identified in chapter IV of the report remain current issues in 2011:

- Exams should be intimately connected with inspection. The Committee members might be disappointed to discover that by tying school accountability to national testing and the use of examination results in league tables, a considerable number of additional issues have emerged which are dominating educational debate a century later.
- The multiplicity of exams should be reduced. In 1911, this could be described as more of a practical problem, arising from the development of geographical regions and the existence of many separate qualifications for entry to different professions. However, the need to provide school accountability has proved to create its own problem of a multiplicity of national tests. Added to this is the debate surrounding the multiple purposes to which the results from examinations are put; a twist to the 1911 debate which has arisen as a consequence of making fewer examinations serve more purposes.
- External exams should be focussed on a clear purpose of helping schools to provide a broad education to age 16, which would provide the foundation for a variety of future study.

The recommendations of the 1911 report led to the School Certificate Examination system, and a more structured curriculum, as described in 'Differentiation of the Curriculum for Boys and Girls Respectively in Secondary Schools'. (1923):

> *These Regulations provide that the minimum curriculum for pupils between the ages of 12 and 16 must include English Subjects, Foreign Languages, Mathematics, Natural Science and Art. We understand that the existing practice is to require the continued study of History, English, a foreign language, Mathematics and a branch of Natural Science throughout this stage, with individual exceptions – general exceptions being allowed only on special grounds*.

The School Certificate required students to pass five subjects, including a humanity, language and maths/science (Watts, 2008). This system has been echoed very recently:

> *So we will introduce a new award – the English Baccalaureate – for any student who secures good GCSE or iGCSE passes in English, mathematics, the sciences, a modern or ancient foreign language and a humanity such as history or geography.* (The Importance of Teaching – The Schools White Paper, 2010)

Summary

The purpose of this article has been to mark the centenary of the 1911 document with an overview of the key issues raised and a relatively brief examination of the extent to which they are current today. However, the wealth of detail in the reports examined, much of which has been beyond the scope of the current article to report, is fascinating and it is useful to consider the value of looking back at the thinking behind earlier educational decisions. It is easy to think of our twenty-first century selves as sophisticated, critical thinkers and to assume that our predecessors a century ago must have been less well-versed, or more simply equipped, or just led a different life with fewer issues. Close acquaintance with the detail in the 1911 report suggests far otherwise. The Committee did not consist of educational philanthropists making comments from an ivory tower of prestige or privilege. Rather, it was made up from experienced educationalists, with practical experience of conditions in schools who backed up their recommendations with practical examples. The paper is studded with evidence from relevant sources. The 1911 document, and many other similar documents (for example, the six Hadow reports published between 1923 and 1933), are extremely detailed and set out

very clearly the thinking behind the decisions that were made. Tracing the outcomes of those decisions, through the legislation which followed, and into policy and practice can inform current educational debates, particularly in instances where consideration is being made of similar initiatives to those which have gone before. It is in these instances that it is possible to be informed by hindsight.

References

Aldrich, R. & Gordon, P. (1989). *Dictionary of British Educationalists*. London: Woburn.

Apprenticeship, Skills, Children and Learning Act, 2009. (c.22). London: HMSO.

Board of Education (1911). *Report of the Consultative Committee on Examinations in Secondary Schools*. London: HMSO.

Children, Schools and Families Committee (2008). 'Testing and Assessment' HC (2007–08) 169–1.

UK Parliamentary Services (2009). Hansard (1803–2005). Available at http://hansard.millbanksystems.com accessed on 14.2.2011.

Hartley, C. (Ed.) (2003). *A Historical Dictionary of British Women*. London: Europa Publications Ltd.

Report of the Consultative Committee (1923). *Differentiation of the Curriculum for Boys and Girls Respectively in Secondary Schools*. Reproduced in D. Gillard, (2006), *The Hadow Reports: an introduction*. www.educationengland.org.uk/articles/24hadow.html

The Conservative Manifesto (2010). Available at http://media.conservatives.s3.amazonaws.com/manifesto/cpmanifesto2010_lowres.pdf accessed on 14.2.2011

The Importance of Teaching – The Schools White Paper, 2010. (CM 7980). London: HMSO.

Transactions of the Royal Historical Society (1945). *Back Matter*. Fourth Series, 27, 97–106. Published by: Royal Historical Society.

Watts, A. (2008). Cambridge Local Examinations 1858–1945. In: S. Raban (Ed.), *Examining the World. A History of the University of Cambridge Local Examinations Syndicate*. Cambridge University Press: Cambridge.

Evaluating Senior Examiners' use of Item Level Data

Hannah Shiell and Nicholas Raikes Research Division

Many of CIE and OCR's written examination scripts are now scanned and marked on screen by examiners working on computers. One benefit arising from on-screen marking is that the marks are captured at item or question-part level and are available for analysis in Cambridge within hours of being submitted by examiners. Cambridge Assessment now routinely analyses these item marks and provides subject staff and senior examiners with reports containing Item Level Data (ILD) for nearly all examinations marked on screen. In this article we present findings from an evaluation of senior CIE and OCR examiners' use of these Item Level Data reports.

Background

Historically, CIE and OCR's written examinations were marked on paper and usually only the total marks were captured electronically. Consequently, if item marks were to be analysed they first had to be keyed in from a sample of written scripts, and this constrained the availability of item level data. With the introduction of on-screen marking, however, marks are now routinely captured at item level for a large and growing number of CIE and OCR's written examinations.

In addition to introducing on-screen marking, Cambridge Assessment has made a major investment in infrastructure to provide research and evaluation staff with:

- a data warehouse providing easy access to operational data, including item marks;
- statistical analysis and reporting tools;
- automation tools (for automating and scheduling analysis and reports);
- an Intranet Portal for publishing statistical reports and data to colleagues across the organisation.

This new infrastructure has enabled us to start routinely producing ILD reports for most CIE and OCR examinations marked on screen. An indication of the scale of this activity is that during peak periods last summer (2010) we analysed 60 million marks per night across nearly 600 examinations.

The nature of the Item Level Data provided

Previous work in Cambridge Assessment identified the kinds of Item Level Data and presentation most useful to subject staff and senior examiners (Johnson, Gill, Elliot and Black, 2006).

We now produce ILD reports on two occasions: firstly during marking, then again after grade boundary marks have been set and candidates' grades are known. The first set of reports are provided to assist subject staff and senior examiners with tasks relating to the current examination, such as providing reports on the candidature's performance and recommending grade threshold marks. The second set of ILD reports, provided once marks have been finalised and candidates' grades determined, are to assist with post-hoc evaluations of the examinations to help identify any improvements that can be made in future examinations. ILD reports are made available as web pages on our Intranet Portal and as documents in pdf format. Few senior examiners

have access to our Intranet, so electronic copies of the pdf reports are sent to them; they may also be shown ILD when attending meetings at our offices.

The following types of output are produced during marking (all updated nightly):

- item statistics (omit rate, facility overall and by quartile, correlation between item marks and overall marks excluding the item);
- item curves (plots of facility by quartile);
- item mark distributions;
- overall internal reliability (Cronbach's alpha);
- overall mark distribution and summary statistics (mean, standard deviation, minimum and maximum mark, all presented overall and by quartile).

OCR generally sends all this information to senior examiners, but CIE initially only sends the item statistics, supplying other information on request since it can amount to many pages of output.

Similar output is produced once candidate grades are known, but this time including grade distributions and breaking information down by grade, sex and, for CIE, country. Sample output can be seen in Figure 1, a screenshot of the item curve and mark distribution chart for one item for one CIE country (details identifying the county and examination have been redacted).

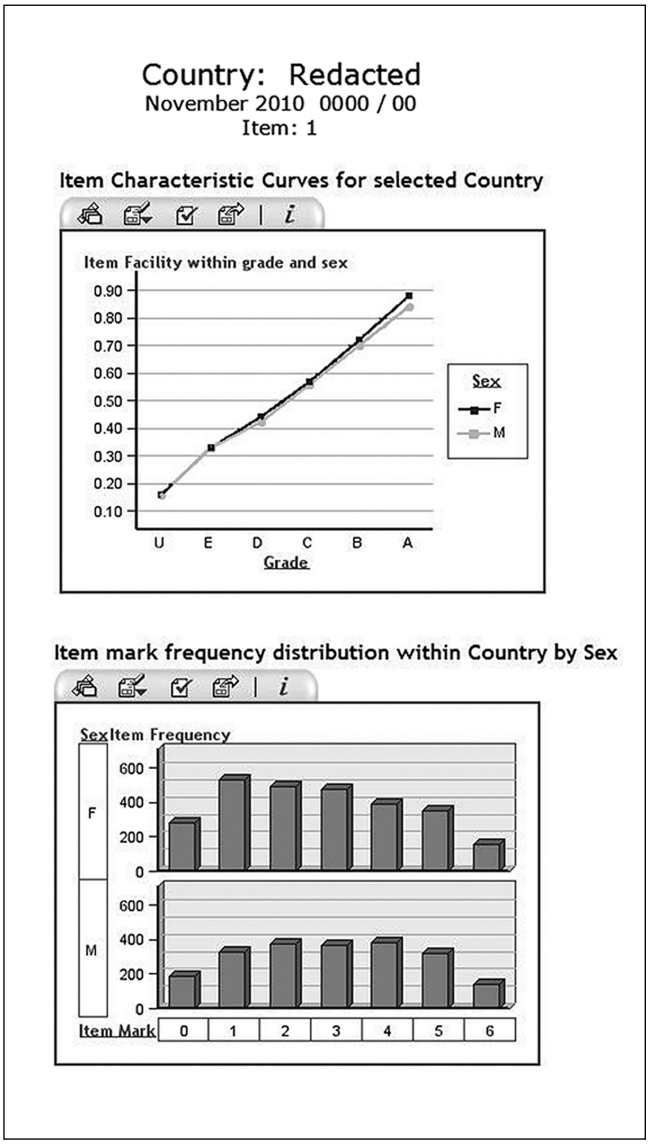


Figure 1: Sample CIE output produced once candidate grades are known

Benefits Review

As part of a wider review of the benefits realised from routinely producing Item Level Data, we solicited feedback from senior examiners.

A short online questionnaire was developed following discussion with subject staff responsible for working with senior examiners using ILD. The questionnaire was reviewed by a panel of researchers not previously involved in the study prior to being used.

We emailed the following senior examiners and invited them to complete the online questionnaire about their use of ILD:

- Principal Examiners and Setters of examinations marked on screen in June 2010 (OCR) and November 2010 (CIE).
  NB: in many cases the same individual was both setter and principal examiner, i.e. he or she both set the question paper and led the marking. The roles are not necessarily combined, however.

Questionnaire findings

The response rate was 71% for CIE (58 responses from 82 invitations) and 59% for OCR (159 responses from 269 invitations).

Some 86% of CIE respondents and 82% of OCR respondents reported that they used ILD.

When asked to assess how helpful they found ILD overall, 78% of CIE respondents and 79% of OCR respondents reported that they found it helpful or very helpful (ratings were made on a five point scale: 2 = very helpful, -2 = very unhelpful). The actual numbers of respondents in each category are shown for CIE in Figure 2 and for OCR in Figure 3.

The senior examiners were also asked to provide feedback on specific uses of ILD. The majority of respondents from both CIE and OCR found ILD helpful or very helpful:

- when writing reports to teachers on candidate performance (see Figure 4 and Figure 5 for the exact number of respondents from CIE and OCR in each category);
- when filling in their 'SRS forms' (Figure 6 and Figure 7). These are the forms on which Principal Examiners make their initial recommendations on where grade boundary marks should be set;
- when identifying items which were harder or easier than expected (Figure 8 and Figure 9), or which did not discriminate as expected between candidates of different 'ability' (as indicated by candidates' total marks) – see Figure 10 and Figure 11.

When asked whether they felt adequately supported in their use of ILD, 76% of CIE respondents and 74% of OCR respondents answered "yes". Given that ILD are relatively novel to many of our senior examiners, this finding is encouraging, though clearly there is scope to improve the support provided. Our current support centres on written documentation explaining each part of the ILD, presentations at meetings, and individual support from subject staff. Improvements suggested by respondents included provision of a separate quick reference glossary of the statistical terms, together with additional written documentation giving examples of use.

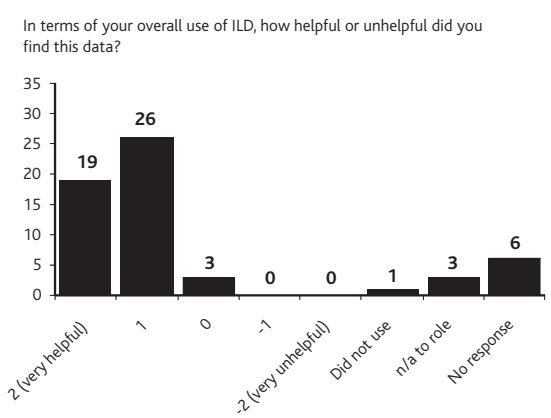


Figure 2: Overall usefulness of ILD – CIE Respondents

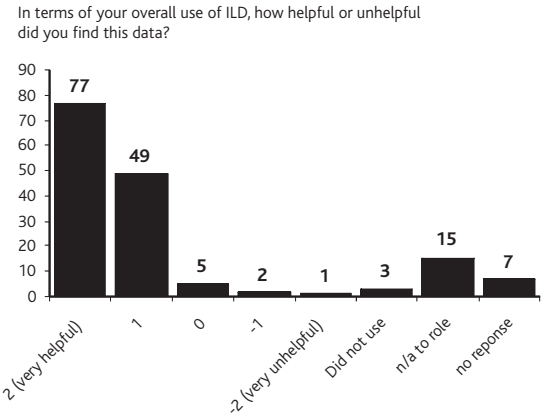


Figure 3: Overall usefulness of ILD – OCR Respondents

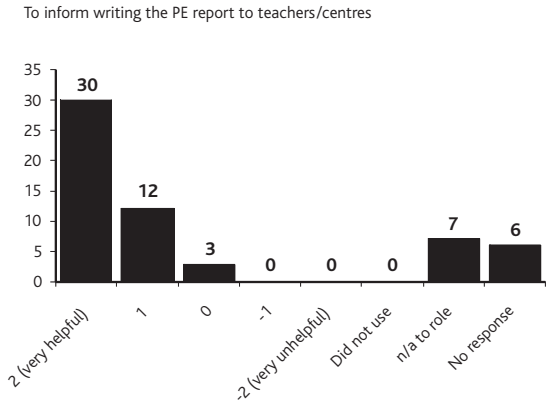


Figure 4: Use of ILD when writing reports to teachers/centres – CIE respondents

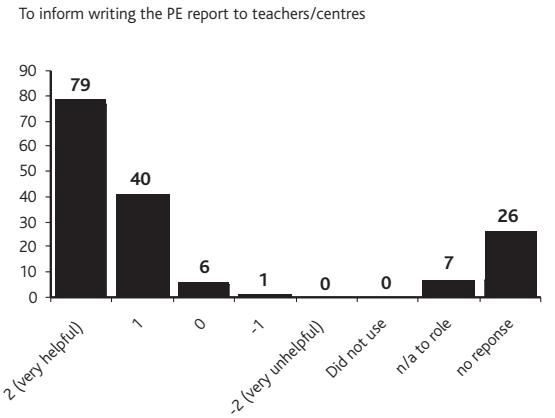


Figure 5: Use of ILD when writing reports to teachers/centres – OCR respondents

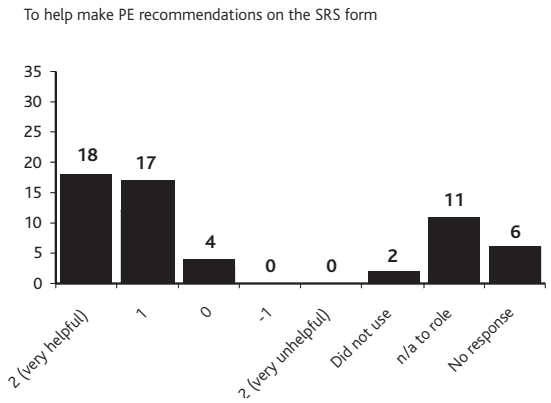


Figure 6: Use of ILD when completing the SRS form relating to recommending grade boundary marks – CIE respondents

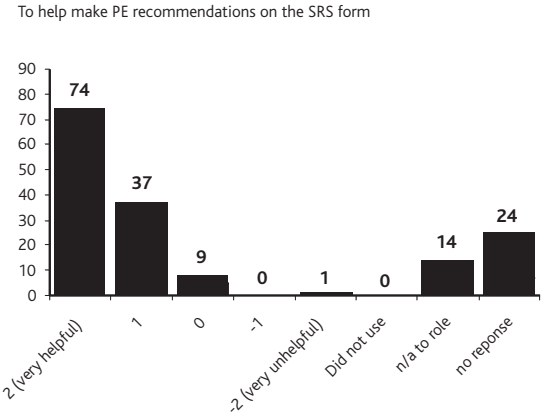


Figure 7: Use of ILD when completing the SRS form relating to recommending grade boundary marks – OCR respondents

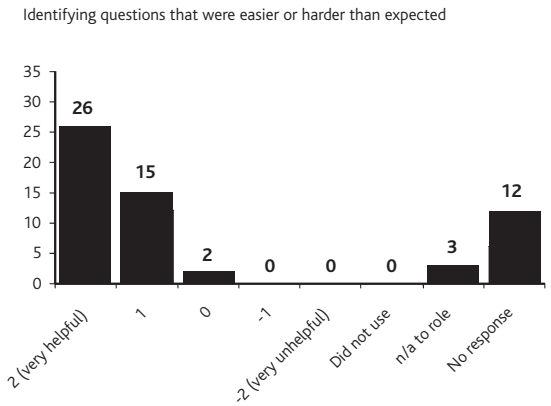


Figure 8: Use of ILD for investigating question difficulty – CIE respondents

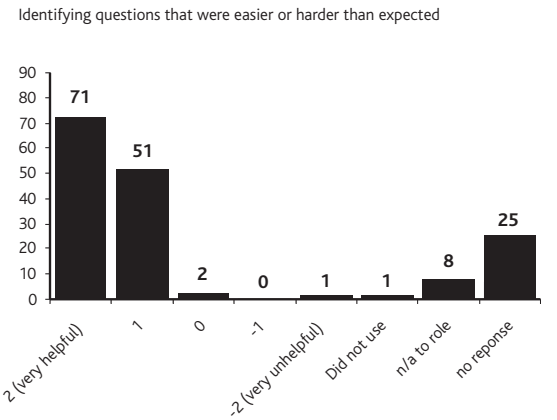


Figure 9: Use of ILD for investigating question difficulty – OCR respondents

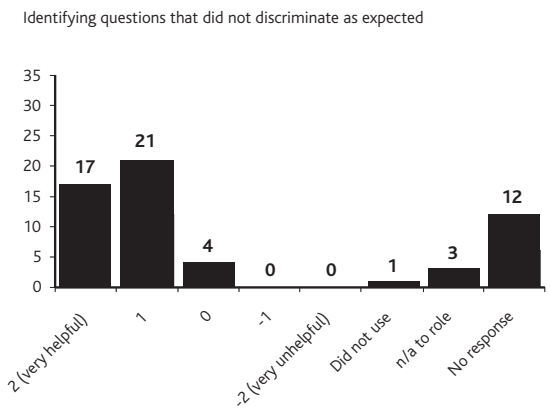


Figure 10: Use of ILD for investigating question discrimination – CIE respondents

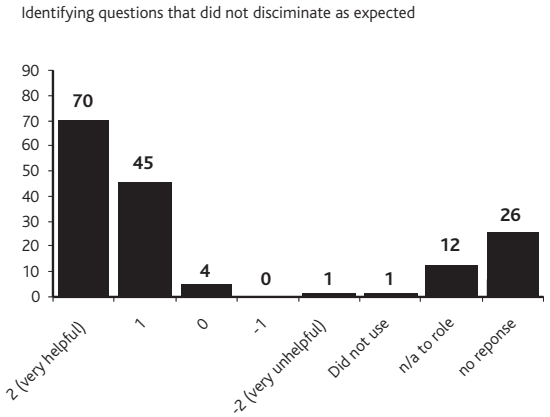


Figure 11: Use of ILD for investigating question discrimination – OCR respondents

Conclusion

The questionnaire findings provide evidence that Cambridge Assessment has successfully introduced routine reporting of Item Level Data to senior CIE and OCR examiners, and that the reports provide helpful information that is widely used. Further work would be required to probe exactly how the information is used. The main limitations of the study are those which generally affect questionnaire-based studies, principally an unquantifiable self-selection bias arising from examiners deciding whether to complete the questionnaire, and the degree to which participants were willing to be open with us and provide accurate and complete answers.

Acknowledgement

We would like to thank Jo Ireland for her help with running the online questionnaire, and all the senior examiners who so generously gave up their time to complete it.

Reference

Johnson, N., Gill, T., Elliott, G. and Black, B. (2006). *Opportunities for evaluating assessments using item level data. Report on 2005–06*. Unpublished internal Cambridge Assessment report.

EXAMINATIONS RESEARCH

Practical issues in early implementation of the Diploma Principal Learning

Victoria Crisp and Sylvia Green Research Division

This short article reports on some of the findings from an interview study conducted in the first year of implementation of the 14–19 Diplomas. The Diplomas were introduced by the Labour government as part of wider educational reforms (DfES, 2005a, 2005b). They were designed to prepare young people for the world of work or for independent study and are intended to combine theoretical and applied learning, to provide different ways of learning, to encourage students to develop skills valued by employers and universities, and provide opportunities for students to apply skills to work situations in realistic contexts. They are also intended to contribute to ensuring that a wide range of appropriate learning pathways are available to young people, thus facilitating increased participation and attainment. The Diplomas are available at Levels 1, 2 and 3 and rather than being taught by an individual school or college, they are available through consortia consisting of a small group of schools and/or colleges working collaboratively.

The Diploma is a composite qualification which is made up of the following elements: principal learning; generic learning; additional and specialist learning. The current research focused on the Principal Learning (PL). The Principal Learning components are specific to a domain or 'line of learning'. Learning through experience of simulated or real work contexts, through applying and practically developing skills, as well as theoretical learning, is emphasised. The PL components are assessed predominantly via assignments which are internally marked and externally moderated. Teaching of Diplomas in the first five 'lines of learning' began in September 2008 with a further five beginning in September 2009 and four in September 2010.

Several initial evaluations of Diploma implementation and other sources have already provided some insights on various issues. One publicly prominent point has been that the uptake of the Diploma was initially lower than expected. The uptake of any course is likely to be strongly affected by whether learners and teachers have a good understanding of that course in order to make informed choices. McCrum *et al.* (2009) interviewed Year 11 students and found that many had limited or incorrect knowledge about Diplomas and that it tended to be

seen as narrowing their options for the future. It seems that early in the implementation of the Diploma advice and guidance on the Diploma was only being offered to students who expressed an interest (Ofsted, 2009).

Another issue raised is that of whether the Diploma provides adequate preparation for higher education study. There was less involvement of HEs in the development of the Diplomas than employers and a survey of representatives involved in the development process suggested that there was slightly lower confidence of the Diploma meeting the needs of students for higher education study (Ertl *et al.*, 2009). However, many higher education institutes are accepting at least some Diploma lines of learning for relevant courses and the broadening of learning styles encouraged within Diploma learning is in keeping with developments in higher education to refine the nature of undergraduate learning (Richardson and Haynes, 2009).

There was some evidence of concern that some students might be unable to pass functional skills and thus would not complete the Diploma. This led some schools to set entry requirements for Level 2 Diplomas based on Key Stage 3 achievements (O'Donnell *et al*., 2009).

Further issues identified relate to practicalities of the Diplomas. For example, some problems with collaboration between schools or colleges, or a reluctance to collaborate as consortia has been found in some cases (AoC, 2009). Designing compatible timetabling was challenging with some clashes occurring (AoC, 2009; Ofsted, 2009). Also, whilst travel between sites was in some cases not problematic due to good public transport and relatively close sites, for some consortia there were challenges in this regard (Ofsted, 2009). During preparation for the Diploma, funding was available to assist in its introduction. However, some colleges considered the method of government funding overly bureaucratic requiring negotiations between schools/colleges in a consortium (AoC, 2009).

Thus, a range of challenges around the Diploma had emerged in early implementation. This study aimed to further explore such themes.

Method

Six consortia running Phase 1 Diplomas in the first year of implementation agreed to take part in this research. These groups of schools/colleges were running the Diploma in Creative and Media, IT, Engineering or Society, Health and Development. All were teaching Level 2 Diplomas, plus two consortia were running either the Level 1 or 3 Diploma in addition. The consortia were visited and, at each, one or more teachers and (in all but one case) a number of learners were interviewed. Learners were interviewed in pairs or groups of three. In total, 11 teachers and 27 learners were interviewed. The visits were made in March to May 2009, thus, the insights gathered are from towards the end of the first year of teaching. The interviews were semi-structured in nature and covered a range of themes relating to the assessments, the learning occurring and various practicalities. The current article will report on the latter. The interview data were analysed by grouping comments by theme and summarising the views expressed.

Findings

The summarised views on the themes relating to practicalities are presented below.

Logistics of moving between sites

Teachers' views

At two of the centres visited, there was no sharing of learners between schools in the consortia and hence no additional travel involved, except for organised induction days or visits to businesses. At these centres teachers were keen to minimise travel, so as to avoid any associated difficulties. At another consortium, learners were taught in two separate centres but the locations of these were linked by playing fields. At the three remaining consortia, learners spent one or two complete days a week at a centre other than their home school. This was generally not problematic because the second centre was not far away, bus services were available and in at least one case these buses were free. Movement between sites was felt to be more manageable when they were dealing with whole days. There was some indication of teachers being more willing for Level 3 learners to study across sites than Level 2 learners. Two teachers mentioned that they had heard of problems relating to travel from other consortia or other schools within their consortia. Difficulties included the cost of travel, taxis not turning up and students arriving late. Transport did seem to be a more general concern for some centres, particularly where sharing of students was likely to increase over time. An associated difficulty mentioned by one teacher was that differences in behaviour policies between centres made it harder for him to apply sanctions for poor behaviour.

Learners' views

Those students whose Diploma learning was based in one place, apart from occasional induction days or trips, reported no problems in relation to transport. At the consortium where students move between two linked sites this was usually unproblematic but inclement weather could make it difficult to walk across the playing fields. This would mean a longer walk between centres or, if a driver was available, a minibus might be organised. At another consortium, students reported that travelling to their second place of learning for their Diploma was unproblematic due to a convenient bus service. At a further consortium, learners would soon be travelling to an additional centre for some classes requiring two bus journeys. This was a worry for some students due to cost and a lack of financial assistance.

Deadlines, scheduling and timetabling (how deadlines fit in with other parts of the Diploma and other courses)

Teachers' views

Generally, no major problems were reported by the teachers in relation to deadlines and scheduling, although some noted that it was still early days. In most cases students were thought to be coping with the demands of work for different aspects of their Diploma and for other courses. Some teachers sensed a degree of tension for students as deadlines approached, but the Diploma was not thought to have added to the pressure and teachers tried to prevent problems by making sure there was time to complete work in lessons. Where common timetables had been agreed between centres sharing learners, this worked well. However, in some cases there were reports of a degree of tension between centres over what should take precedence. There were examples of clashes between classes and with events at the home school leading to missed lessons and learners needing to catch up. At one consortium agreeing between centres on the scheduling of functional skills tests and on who was responsible for paying for them was problematic. One teacher commented that co-ordinating classes between two centres had

been relatively easy this year, but that co-ordinating between more centres in future years would be more difficult in terms of covering linked topics in parallel.

Learners' views

Most learners reported that, so far, they had been able to meet their deadlines for completing assignments and that the deadlines set were realistic. So far, work for different strands of the Diploma and for other courses had reportedly fitted together without problems. However, some felt that it was sometimes challenging to keep on top of their workload and that dealing with their work for different subjects was sometimes hectic. Deadlines were generally viewed positively because they were felt to help them learn more (e.g. report writing skills, managing their own work) and motivated them. Some students commented that they were keen not to fall behind with work as it would be difficult to catch up. Learners reported finishing off work at lunchtimes, after school in supervised sessions, or at home. In two consortia there were some timetabling difficulties which had resulted in some missed lessons that learners had to catch up. One student commented that because their teachers did not know what to expect in the first year of the course this had led to some initial difficulties with timetabling, for example.

Funding

Teachers' views

There was concern about funding for equipment and materials to support Diploma teaching. At one school the funding for the course had arrived at the school, but had been delayed in arriving in the relevant departmental budget. Several teachers reported that funds had been available for the first year but were concerned that this might not continue.

Workload

Teachers' views

One teacher commented that the PL required "*an awful lot of work*", perhaps implying that the amount of work for the assessments was excessive in her view. This linked to a comment from another teacher at the same consortium, who felt that insufficient curriculum time had been allocated to the Diploma at her school.

Policy issues and fast introduction of the Diploma

Teachers' views

Several teachers noted issues around the newness of the qualification. One suggested that a longer pilot period would have been valuable, another that it would take time to find the best ways of delivering the course for their students and two others commented that more guidance (e.g. exemplar work, training) would have been helpful. More exemplar work and written guidance was likely to be available in the second year of delivery.

A number of comments related to policy. One teacher felt that schools and colleges had not been adequately consulted with regard to the Diploma and that policy decisions had not been guided by experience and education. Another thought that a complete change to replace GCSEs altogether with Diplomas would be easier for students to understand, and that the current situation left learners somewhat confused as to the relationship between their different courses. A teacher who was very keen on the Diplomas in her subject area was not in favour of the introduction of Diplomas in 'academic' subjects as existing qualifications fulfil these goals sufficiently. An FE teacher for Creative and Media commented that they already ran the National Diploma at level 3 in their college, and that this meant it was not financially viable to run the new 14–19 Diploma at level 3 as well.

Composite nature of the qualification

Teachers' views

Teachers expressed concern that some students were struggling with maths functional skills which could mean failing the Diploma overall.

Learners' views

One pair of learners expressed concerns about aspects of the Diploma qualification. They were worried about what would happen if they failed one section of the assessment and whether this would mean an overall fail or whether retakes would be possible. They also expressed concerns about recognition of the Diploma by universities.

Discussion

Whilst this research was small-scale, it provides further insights into practical issues in the early days of implementation of the 14–19 Diplomas. Of those consortia where students were studying for their Diploma across more than one site, there were a few difficulties noted in terms of moving between sites and timetabling clashes. This echoes such logistical issues identified in some consortia by earlier research/evaluation (Ofsted, 2009; AoC, 2009). Other consortia had planned compatible timetables across sites, organised classes into whole days spent at one site (rather than moving between sites part way through a day) and were fortunate in terms of public transport links such that these practicalities were unproblematic. Funding was also raised as a constraint. For some consortia, the funding provided in the first year had been very beneficial. For others there had been issues with the funding arriving at a centre, but taking some time to become designated to the appropriate budget to assist with resources specific to the Diploma. This is likely to be a 'teething problem' at the local level which should hopefully be avoided in future. Another concern related to the longevity of funding, with some worries that funding may not continue in the future at the current level. The AoC (2009) noted that due to initial low uptake some colleges were currently subsidising the implementation of Diploma courses, and that this would not be sustainable long term. Some difficulties in relation to functional skills were noted, specifically in relation to collaborating over timetabling and prioritisation and issues around who is responsible for examination fees.

Whilst some of the challenges experienced by some consortia may have been short-term 'teething' problems which may now have been resolved, some may be longer term issues or may become more problematic as numbers of students or collaboration between centres increases. In contrast to some of the practical difficulties sometimes experienced in early Diploma teaching and learning, other themes explored in the interviews (to be reported in full elsewhere) suggested substantial positive feeling amongst teachers and learners about the aims of the Diploma, and the nature of the learning encouraged. A wide range of subject specific and wider skills, that would be valued in work places (e.g. independent working, project management, teamwork and interpersonal skills, research, report writing), were reportedly being developed via Diploma courses. Most of the teachers were enthusiastic and most learners were motivated by the work.

References

AoC (2009). *AoC Diploma Survey Report – Results, analysis, conclusions and recommendations*. London: Association of Colleges. Available at: http://www.aoc.co.uk/en/newsroom/aoc_news_releases.cfm/id/C5B28C9E-78EC-4E75-BAE3D47ECDF3BD48

DfES (2005a). *14–19 Education and Skills White Paper*. London: DfES. Available at: http://publications.dcsf.gov.uk/eOrderingDownload/CM%206476.pdf

DfES (2005b). *14–19 Education and skills implementation plan*. London: DfES. Available at: http://publications.teachernet.gov.uk/eOrderingDownload/2037-2005PDF-EN-01.pdf

Ertl, H., Stanley, J., Huddleston, P., Stasz, C., Laczik, A., & Hayward, G. (2009). *Reviewing diploma development: evaluation of the design of the diploma qualifications*. London: Department for Children, Schools and Families. Available at: http://www.dcsf.gov.uk/research/data/uploadfiles/DCSF-RW080%20(Rev).pdf

McCrum, E., Macfadyen, T., Fuller, C., & Kempe, A. (2009). *Vocational education and training: some perspectives from Year 11*. Paper presented at the British Educational Research Association Annual Conference, 3–6 September 2009, Manchester.

O'Donnell, L., Lynch, S., Wade, P., Featherstone, G., Shuayh, M., Golden, S., & Haynes, G. (2009). *National evaluation of Diplomas: Preparation for 2008 delivery*. London: Department for Children, Schools and Families. Available at: http://publications.dcsf.gov.uk/eOrderingDownload/DCSF-RW079.pdf

Ofsted (2009). *Implementation of 14–19 reforms, including the introduction of Diplomas*. London: Office for Standards in Education, Children's Services and Skills.

Richardson, W. & Haynes, G. (2009). *National evaluation of diplomas – findings from the 2008 survey of Higher Education Institutions on their implementation and impact*. London: Department for Children, Schools and Families. Available at: http://www.dcsf.gov.uk/research/data/uploadfiles/DCSF-RR145.pdf

The effect of changing component grade boundaries on the assessment outcome in GCSEs and A levels

Tom Bramley and Vikas Dhawan Research Division

Acknowledgements

This paper is based on one section of a report* commissioned by Ofqual (the exams regulator in England) as part of its Reliability Programme. For more details on this programme, see http://www.ofqual.gov.uk/research-and-statistics/92-articles/20-reliability. We would like to thank Ofqual and its technical advisory group for their feedback. The opinions expressed are those of the authors. Figure 1 and its commentary were not part of the original report.

Introduction

Investigations of assessment reliability are concerned with answering the question 'how would the assessment outcomes change if the assessment were replicated?' The answer to this question depends on what factors are held constant and what factors change on replication. For example, the examination questions could be different, or the markers (examiners) could be different – or both these could be held constant and the only change might be in the mood or level of preparation or other factors internal to the examinees. A further factor relevant to GCSE and A level assessments is that these are graded examinations, where grade boundaries are set on the raw mark scale of each of the units/components comprising the assessment. These boundaries are then aggregated in a particular way depending on the type of assessment to produce the overall grades for the assessment. It is therefore possible to consider a replication scenario where questions, markers and examinee internal factors remain the same, but the grade boundaries (and hence the grade outcomes) are different.

A variety of sources of evidence can be used to inform the decisions about where to set the grade boundaries, including:

- 'archive' scripts at the key grade boundary marks from previous sessions;
- information about the size and composition (e.g. type of school attended) of the cohort of examinees;
- teachers' forecast grades;
- the distribution of scores (mean, SD, cumulative % of examinees at each mark);
- at GCE, 'putative' grade distributions (grade distributions generated by matching examinees with their GCSE results and taking account of changes in the 'ability' of the cohort of examinees from a previous[1] session, as indicated by changes in the distribution of mean GCSE scores;
- experts' judgements about the quality of work evident in a small sample of scripts covering a range of consecutive marks (total scores) around where the boundary under consideration is expected to be found;
- experts' judgements about the difficulty of the question paper;
- other external evidence suggesting that the particular unit/component (or assessment as a whole) had previously been severely or leniently graded and needs to be 'brought into line' with other examination boards, or with other similar subjects or specifications within the same board.

1. Usually this is the previous session with a cohort believed to be most similar to the current session's cohort, e.g. for a June 2009 unit, the June 2008 session might be used rather than the January 2009 session.

These pieces of evidence do not necessarily always 'point in the same direction', and therefore they need to be weighed appropriately – a matter which ultimately requires human judgement, although it is fair to say that most weight is given to statistical methods that take account of changes in the 'ability' of the cohort. Given that it is therefore not possible to determine exactly what the grade boundaries 'should' be, it is of interest to investigate what the impact of slightly different decisions at unit/component level would be on the grade distributions at whole assessment level. In particular, it seems likely that the evidence for any particular grade boundary decision could support two possible boundary marks, and perhaps more.

Whilst it would in principle be possible to carry out an actual replication of the grade boundary setting process, varying some of its characteristics (e.g. decision-making personnel, scripts viewed etc.), considerable if not prohibitive logistical (and financial) problems would arise.

Therefore, we carried out a simple 'sensitivity analysis' in order to determine the effect on assessment grade boundaries of varying the (judgementally set) key grade boundaries on the units/components by ±1 mark. In this paper we report the results of this analysis for two assessments with different structures – a tiered 'linear' GCSE, and a 6-unit 'modular' A level. The data came from the June 2009 examination session administered by OCR.

The effect of varying component grade boundaries on a tiered, linear GCSE examination

In this assessment, Foundation Tier examinees took two written papers and a coursework component. Higher Tier examinees took two different written papers and the same coursework component (which therefore had the same grade boundaries for each tier).

In linear assessments, there are two ways of deriving the aggregate grade boundary from the component grade boundaries. The first, known as 'Indicator 1', is the simple aggregate of the component grade boundaries, taking account of the weight of each component in the aggregate total. In this GCSE the two written papers each carried 40% weight and the coursework 20%, and their paper totals were in these proportions, which meant that indicator 1 could simply be obtained by adding up the grade boundary marks on the three components. Tables 1 and 2 below show the range of possible boundaries at grade C (Foundation) and grade A (Higher) obtainable if the boundaries on some or all of the three components were changed by ±1 mark. The column '# combinations' shows how many ways there were of arriving at that particular aggregate boundary mark. Clearly there is only one way of arriving at a mark 3 lower or higher – that is, by raising or lowering each component boundary by 1 mark. However, the various other permutations lead to more ways of arriving at boundaries within this range.

Tables 1 and 2 show that the possible values for the actual aggregate grade boundary could have led to fluctuations covering a range of up to 9.5 percentage points in the pass rate at grade C on the Foundation Tier, and up to 6 percentage points in the pass rate at grade A on the Higher Tier. Even a ±1 mark difference from the actual boundary would have given a range of ≈3 percentage points at grade C on the Foundation Tier and ≈2 percentage points at grade A on the Higher Tier.

Table 1: Foundation Tier – possible aggregate grade C boundaries (Indicator 1 only)

Aggregate boundary	*# combinations*	*Cumulative % of examinees at Grade C*
93	1	34.50
92	3	35.96
91	6	37.74
90	7	39.47
89	6	41.04
88	3	42.50
87	1	43.96

Table 2: Higher Tier – possible aggregate grade A boundaries (Indicator 1 only)

Aggregate boundary	*# combinations*	*Cumulative % of examinees at Grade A*
167	1	13.71
166	3	14.47
165	6	15.41
164	7	16.55
163	6	17.30
162	3	18.47
161	1	19.70

The second method of calculating the aggregate boundary, known as 'indicator 2', involves finding the mark on the aggregate distribution of marks (the distribution obtained by adding together each examinee's mark on each component, appropriately weighted) where the cumulative percentage of examinees obtaining that mark corresponds most closely to the percentage obtained by taking a weighted average of the cumulative percentage of examinees at that particular boundary on each of the components. Indicator 2 is usually closer to the mean aggregate mark than indicator 1, which means it is usually lower than indicator 1 at the higher boundaries, and vice versa. The Code of Practice (Ofqual, 2009) allows the awarding panel to choose any mark between (and including) the two indicators as the final aggregate boundary mark. The default position is to take the lower of the two indicators[2].

The effect of including indicator 2 was to increase the range of possible boundaries by one mark down to a mark of 86 at grade C on the Foundation Tier (cf. 87 with indicator 1, see Table 1). The effect was greater at grade A on the Higher Tier, where it increased the range of possibilities by a further six marks down to a possible mark of 155 (cf. 161 in Table 2).

In statistical tables of examination results, the outcomes for the two tiers of the examination are combined rather than published separately. Grade A is only available on the Higher Tier, but grade C is available on both tiers, which dramatically increases the number of overall possible outcomes at grade C if the boundaries on all components on *both* tiers fluctuate by ±1 mark. As we have seen, the more extreme outcomes are less likely to arise because they would require a change in the same

2. The Code of Practice states "Whenever the two indicators do not coincide, the grade boundary should normally be set at the lower of the two indicator marks, unless, in the awarders' judgement, there is good reason, as a result of a review of the statistical and technical evidence, to choose a higher mark within the range spanned by the indicators." Ofqual (2009) p.53.

direction on all components. Table 3 below shows the extreme (widest possible) range and a more plausible range based on the most likely aggregate outcomes for the overall grade A and C on the whole assessment (both tiers combined).

Table 3: Overall possible pass rate outcomes (cumulative % of examinees)

	Cumulative % of examinees	
	Extreme range	*More plausible range*
Grade A	8.6 to 16.5	10.3 to 14.4
Grade C	55.9 to 63.2	58.2 to 61.0

In summary, a range of 3–4 percentage points seems like a reasonable range in which the cumulative percentage outcomes at grades A and C on this linear GCSE might fluctuate. This value is contextualised in the discussion section (see later).

The effect of varying unit grade boundaries on a modular 6-unit A level

GCE AS and A levels are 'modular' or 'unitised' – that is, examinees are assessed in discrete units. Most AS levels consist of 2 units, but some contain 3. Most A levels consist of 4 units, but some contain 6. The A levels include the corresponding AS units, plus further 'A2' units. The A2 units do not form a qualification on their own, unlike the AS units. The number and choice of units depends on the specification (syllabus). Most units are 'available' in examination sessions in January and June. Any exceptions or restrictions are stated in the specification. Examinees would generally take AS units in the first year of a 2-year A level course, and the A2 units in the second year. Units can be re-taken individually: in other words if an examinee wishes to improve their aggregate grade they do not need to re-take every unit in the assessment.

Because of this choice and flexibility in modular assessment schemes, a different method for deriving the aggregate grade boundaries is required. AS and A level units have a 'Uniform Mark Scale' (UMS). The key grade boundaries 'A' and 'E' are set on the raw mark scale for each unit, and these raw marks are converted to fixed boundaries on the UMS. The conversion between raw and uniform marks is linear within the A–E range and extended slightly beyond it – see AQA (2009) and Gray and Shaw (2009) for further details. For 6-unit A levels the aggregate grade A boundary is at 480 UMS marks (out of 600), and for grade E it is at 240 UMS.

In terms of the effect on aggregate outcome of changes to the unit boundaries, it is only reasonable to consider the effect of changes made in a particular examination session. This is because once the unit boundaries have been set, they cannot later be changed. So, when considering the effect of changing the boundaries on all units of a 6-unit A level in June 2009 by ±1 mark, it should be emphasised that the vast majority of examinees would already have taken units in previous sessions – probably the three AS units in January and June 2008, and perhaps one A2 unit in January 2009. Table 4 shows part of the breakdown of numbers of examinees taking units in June 2009.

Table 4 makes it clear that nearly half of the examinees aggregating in June 2009 had just taken two or three A2 units in June 2009. Only 3% had taken all six units in June 2009.

Table 4: A 6-unit A level – number of aggregating examinees taking each unit in June 2009 (total N=11,603). Only combinations with more than 100 examinees are shown

AS units			*A2 units*				
Unit 1	*Unit 2*	*Unit 3*	*Unit 4*	*Unit 5*	*Unit 6*	*N*	*%*
✗	✗	✗	✗	✓	✓	2929	25.24
✗	✗	✗	✓	✗	✓	288	2.48
✗	✗	✗	✓	✓	✓	2348	20.24
✗	✗	✓	✗	✓	✓	657	5.66
✗	✗	✓	✓	✓	✓	500	4.31
✗	✓	✗	✗	✓	✓	239	2.06
✗	✓	✗	✓	✓	✓	736	6.34
✗	✓	✓	✓	✓	✓	327	2.82
✓	✗	✗	✗	✓	✓	738	6.36
✓	✗	✗	✓	✓	✓	476	4.10
✓	✗	✓	✗	✓	✓	405	3.49
✓	✗	✓	✓	✓	✓	283	2.44
✓	✓	✗	✓	✓	✓	317	2.73
✓	✓	✓	✓	✓	✓	352	3.03

Changing the boundaries on all six units by ±1 mark would give 3^6 = 729 possible scenarios. Given the complexity of the computations required to derive the final grade distributions (which involve obtaining unit-level UMS distributions going back several years) it was only feasible to investigate a relevant selection of these possible scenarios. The outcomes are shown in Table 5.

Table 5 shows that the variability of aggregation outcomes at grade A was ≈3 percentage points when the grade A boundaries on the 6 units were moved by ±1 mark. Changing all the AS units simultaneously only affected the outcome by about 0.5 percentage points. Not surprisingly, given the entry patterns shown in Table 4, changes to the A2 units had more impact – changing the boundary on either Unit 4, Unit 5 or Unit 6 had as much impact as changing the boundary on all three AS units. Unit 5 and Unit 6 on the A2 appeared to be more influential than Unit 4, but given that more of the aggregating examinees had taken Unit 5 and Unit 6 in June 2009 this is not surprising.

Table 5: A 6-unit A level – effect of varying June 2009 unit grade A boundaries on overall % of examinees at grade A (actual outcome in bold)

Unit 1 June 2009	*Unit 2 June 2009*	*Unit 3 June 2009*	*Unit 4 June 2009*	*Unit 5 June 2009*	*Unit 6 June 2009*	*Cumulative % aggregate grade A (N=11,603)*
-1	-1	-1	-1	-1	-1	33.41
0	0	0	-1	-1	-1	32.94
0	0	0	0	0	-1	32.35
-1	-1	-1	0	0	0	32.33
0	0	0	0	-1	0	32.31
0	0	0	-1	0	0	32.07
0	**0**	**0**	**0**	**0**	**0**	**31.84**
0	0	0	+1	0	0	31.60
0	0	0	0	0	+1	31.39
+1	+1	+1	0	0	0	31.36
0	0	0	0	+1	0	31.27
0	0	0	+1	+1	+1	30.79
+1	+1	+1	+1	+1	+1	30.35

Discussion

Given all the sources of information that can be used in setting grade boundaries, some of which relate to different definitions of what it might mean to 'maintain a standard' (see, for example, Baird, 2007; Coe, 2010; Newton, 2010) and which therefore can suggest different locations for the grade boundaries, it should be clear that the setting of grade boundaries is not a problem with a clear-cut answer. Therefore, it is perhaps of interest to consider how the outcomes might have been different if different decisions had been taken[3]. The analyses presented here give some indication of what such reporting might look like. Two potentially useful ways of quantifying the potential variability in aggregate outcome are:

- to determine the range of possible aggregate outcomes that could have arisen if all relevant key grade boundary decisions at unit/component level had been 1 mark lower or 1 mark higher;
- to discover the largest change to the aggregate outcome that could have arisen from a 1-mark change in the boundary on a single unit/component.

The most obvious factors affecting the sensitivity of the aggregate outcome to decisions on the individual units/components are: i) the number of units/components to be aggregated – the greater the number the less the effect of changes on any one unit/component; and ii) the percentage of examinees on each mark point at the part of the distribution where the grade boundary lies (on each unit in unitised schemes, but on the aggregate distribution in linear schemes[4]). Units with longer raw mark scales, all things being equal, might be expected to have a lower percentage of examinees on each mark point. The correlation of scores among the units can also be expected to have an effect, with changes to grade boundaries on more highly correlated units/components affecting the aggregate more.

A more subtle point relating to unitised assessments is the effect of potential grade boundary changes to the 'conversion rate' of raw marks to uniform marks. Changes that reduce the distance between the A and the E boundary (i.e. lowering the A boundary and/or raising the E boundary) increase the rate of exchange; and vice versa. So whereas on a linear assessment a change to a component boundary changes the aggregate boundary but does not affect the aggregate totals of any examinees, in a unitised assessment a change to a unit boundary does not affect the aggregate UMS boundary but does affect the unit (and hence the aggregate) UMS total of most of the examinees who took that unit. So on a linear assessment (for example a higher tier GCSE) a change to a component grade A boundary could not affect the cumulative percentage of examinees obtaining aggregate grade C, but on a unitised assessment a change to a unit grade A boundary could conceivably affect the cumulative percentage of examinees obtaining aggregate grade E. Admittedly this effect is likely to be very small for the ±1 mark changes we are talking about. In the case of the 6-unit A level reported here, lowering the grade A boundary by 1 mark on all six units would have resulted in an extra 3 examinees (out of 11,603) obtaining an aggregate grade E. Lowering the A boundary and the E boundary by 1 mark on all six units would have resulted in an extra 30 examinees obtaining an aggregate grade E. Interestingly, lowering the E boundary by 1 mark and raising the A boundary by 1 mark on all six units would have resulted in an extra 38 examinees obtaining aggregate grade E! This illustrates the point that the UMS conversion can have some slightly counter-intuitive effects – but supports the claim that the proportion of examinees affected is likely to be very small.

In unitised assessments it is very difficult to gauge or control the impact of changes at unit level because of the large number of different valid combinations of units, from different examination sessions, that can be aggregated to achieve an overall result at assessment level. Decisions made in a particular examination session cannot have any effect on the UMS scores on units from previous sessions. For the new unitised GCSEs, certificated for the first time in June 2010, 'terminal rules' specify that a certain proportion of the units must be taken in the same session that aggregation will take place, which will presumably mitigate this problem to some extent.

We therefore would argue that an appropriate way to quantify 'grading reliability' would be to consider the range of possible outcomes (grade distributions) that could have been obtained if grade boundary decisions taken in a particular session had been slightly different. We have chosen to define 'slightly different' as 'varying by ±1 mark', that is, the smallest difference possible. There would be some justification for taking a wider range, given that the 'zone of uncertainty'[5] in expert judgement of script quality usually spans a range wider than ±1 mark. The results presented here could then be seen as lower bounds.

To put the kinds of variability we have found into context, Table 6 shows the cumulative percentage of examinees obtaining grade A from June 2006 to June 2009 in the two assessments discussed above. This table uses the 'final' data on the system, rather than the data available at the time of awarding used in the analyses presented above, so the numbers of examinees do not exactly match.

Table 6: Grade A cumulative percentages and number of examinees, 2006–2009

Qualification		*2006*	*2007*	*2008*	*2009*
Linear GCSE	%	13.6	12.2	12.5	14.6
	N	3323	3977	4764	5244
6-unit A level	%	28.6	29.9	30.9	31.7
	N	10290	11113	11472	11874

It is very striking how similar the cumulative percentages gaining grade A were from year to year in the period 2006–2009, given that the examinees were different and the size of the entry varied somewhat. In no case was the largest difference between any pair of years more than 3.1 percentage points, and most adjacent pairs of years differed by less than 1 percentage point. On the other hand, the analysis above showed that the possible range of variation in percentage at grade A with *exactly the same examinees* could be from around 2 to 4 percentage points, if boundary marks on all units/components were changed by ± 1 mark. This

3. Of course, examination boards do consider the aggregate effect of the decisions made at unit level at the time when those decisions are taken, in 'modelling' exercises. We are suggesting here that the range of possible fluctuation had decisions been slightly different could be reported more systematically.

4. For linear schemes that use indicator 1 only. If indicator 2 is used then the number of examinees at mark points around the boundary on the individual components is also relevant.

5. The term formerly given to the range of marks over which there was no consensus among a panel of experts that the quality of scripts was definitely worth the higher or lower of two adjacent grades. Nowadays this range is referred to simply as the 'zone' – presumably so as not to give the impression that there is any uncertainty in the process!

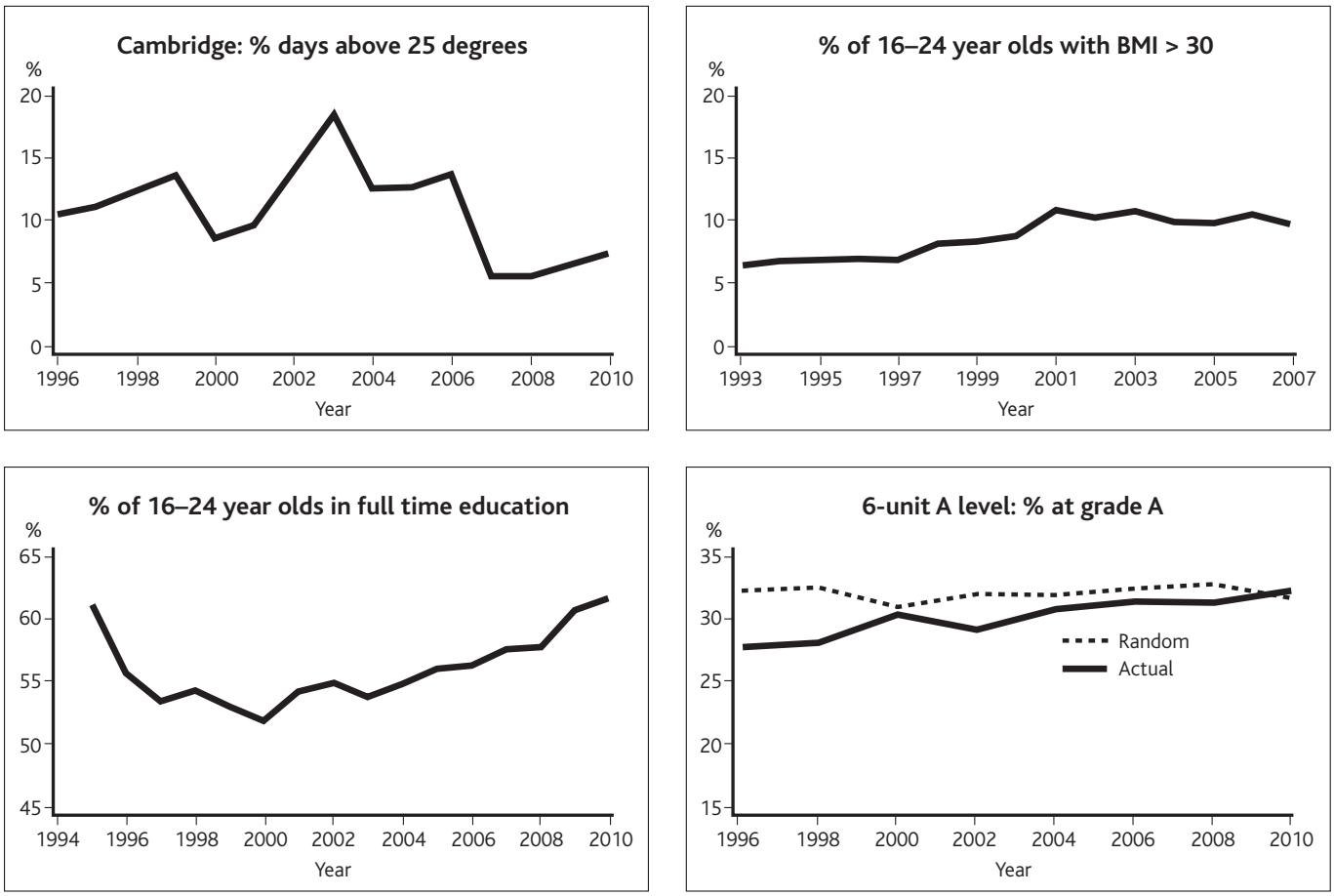


Figure 1: Illustrations of various trends of data expressed as percentages

Data sources for Figure 1

Top left: Cambridge computer laboratory daily weather record http://www.cl.cam.ac.uk/research/dtg/weather/index-daily-text.html Accessed 8/2/11.

Top right: NHS information centre, Body Mass Index (BMI) data. http://www.ic.nhs.uk/webfiles/publications/HSE07/ADULT%20TREND%20TABLES%202007.xls Accessed 8/2/11.

Bottom left: UK National Statistics publication hub: Labour market statistics: educational status, economic activity & inactivity of young people. http://www.statistics.gov.uk/StatBase/xsdataset.asp?vlnk=5740&More=Y Accessed 8/2/11.

Bottom right: Joint Council for Qualifications: inter-awarding body statistics.

suggests that the current statistically driven grade-boundary setting procedures could be 'overfitting' and producing a year-on-year grade distribution that does not fluctuate enough, given all the conceptual conflicts and practical limitations of the standard maintaining process. Of course, given public expectations about 'standards' it might be difficult to explain that a more fluctuating grade distribution is perfectly acceptable. On the other hand, it would help to avoid the pattern that is sometimes seen of steady year-on-year small incremental rises in pass rates that lead to accusations of 'grade drift' (see, for example, Oates, 2009 and its coverage in Paton, 2010).

As an illustration of this point, Figure 1 shows four graphs of time series data where the variable plotted is a percentage. The y-axis covers the same range of percentage points for ease of comparison. It can be seen that while fluctuations in A level % pass rate at grade A in Chemistry are of a similar order of magnitude to the other variables (with the exception of warm days in Cambridge!), there is a tendency for consistent very small increases. By contrast, eight bootstrap samples from the 2009 aggregate distribution in the large-entry 6-unit A level (shown as the dashed 'random' line in the bottom-right graph in Figure 1) showed the kind of fluctuations in pass rate that might be expected if random variation was the only source of year-on-year differences.

The fact that the observed fluctuations are of a similar size to the random fluctuations, but in a more consistent (upward) direction could be explained by saying that in the years when random fluctuations would increase the pass rate, they are the only factor operating, but in the years when they would decrease it, other factors act to cancel them out by more in the opposite direction. However, this does seem rather implausible. A more likely explanation is that awarding panels look for the 'safety in numbers' that the statistical sources of evidence appear to provide, and combine this with a tendency to give examinees the 'benefit of the doubt' when undecided about two adjacent marks for a grade boundary (Stringer, 2008).

In summary, reliability investigations seek to show how outcomes would vary if some factors were changed while others remained constant. One factor affecting outcomes is the decision of the awarding panel on where to locate the grade boundaries on the raw mark scale of each unit/component. Small changes to grade boundaries of the units/components of the linear GCSE and modular A level reported here would have produced fluctuations in the cumulative percentage of examinees reaching the boundary in a 2–4 percentage point range. This is slightly larger than the range of fluctuation that might be expected from random sampling variability (in large entry subjects), and larger than the observed range of changes across a period of several years. We suggest that this finding supports the claim that the observed pass rates do not fluctuate enough in both directions and that the current boundary-setting procedures might be achieving a tighter level of statistical control than is necessary or appropriate.

References

AQA (2009). Uniform marks in GCE, GCSE and Functional Skills exams and points in the Diploma. http://store.aqa.org.uk/over/stat_pdf/UNIFORMMARKS-LEAFLET.PDF Accessed 16/02/10.

Baird, J.-A. (2007). Alternative conceptions of comparability. In: P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education*, **25**, 3, 271–284.

Gray, E., & Shaw, S. (2009). De-mystifying the role of the uniform mark in assessment practice: concepts, confusions and challenges. *Research Matters: A Cambridge Assessment Publication*, **7**, 32–37.

Newton, P.E. (2010). Contrasting conceptions of comparability. *Research Papers in Education*, **25**, 3, 285–292.

Oates, T. (2009). 'Standards are up this year' – what does this mean? The question of standards in public examinations. http://cambridgeassessment.files.wordpress.com/2010/01/the-question-of-standards-in-public-examinations-by-tim-oates1.pdf Accessed 17/5/10.

Ofqual (2009). GCSE, GCE and AEA code of practice, April 2009. http://www.ofqual.gov.uk/files/2009-04-14-code-of-practice.pdf Accessed 08/01/10.

Paton, G. (2010). GCSE and A level results being 'inflated'. *Daily Telegraph*. http://www.telegraph.co.uk/education/educationnews/7528383/GCSE-and-A level-results-being-inflated.html Accessed 17/5/10.

Stringer, N. (2008). *An appropriate role for professional judgement in maintaining standards in English General Qualifications*. Paper presented at the International Association for Educational Assessment annual conference, Cambridge, September 2008.

*Bramley, T. & Dhawan, V. (2010). Estimates of Reliability of Qualifications. Coventry, UK: Ofqual. http://www.ofqual.gov.uk/files/reliability/11-03-16-Estimates-of-Reliability-of-Qualifications.pdf Accessed 26/5/11.

An American university case study approach to predictive validity: Exploring the issues

Stuart Shaw and Clare Bailey CIE

Introduction

Predictive validity research is fundamental to test validation (Davies *et al*., 1999). Predictive validity entails the comparison of test scores with some other measure for the same candidates taken some time after the test has been given (see Anastasi, 1988; Alderson *et al.*, 1995). In psychometric terms, predictive validity is the extent to which a scale predicts scores on some external (future) criterion measure. It is the prediction of criterion performance that is basic to validation. For tests that are used for university selection purposes it is vital to demonstrate predictive validity.

However, establishing predictive validity through relating secondary school performance to later academic performance is fraught with practical difficulties in mounting tracer studies and the problems associated with confounding intervening variables that obscure the effects of another variable (see Banerjee, 2003, for a critique of such approaches to establishing predictive validity). These difficulties notwithstanding, predictive validity is still regarded a vital aspect of the validation process. Moreover, predictive validity research is becoming increasingly necessary as test providers are being challenged to pay greater attention to issues of test comparability – both in terms of the relationships between their own assessment products and those offered by other competitor, examination boards.

A common need for predictive validity is inherent in the process of selecting students for university. Consequently, this article will focus on the research being conducted by University of Cambridge International Exams (hereafter simply 'Cambridge') to ensure that its international assessments prepare students well for continued studies in colleges and universities. The long-term purpose of the research is to highlight the predictive validity of Cambridge assessments and other students' characteristics to predict preparedness for and continued academic success at U.S. universities in terms of first year Grade Point Average (GPA).

This study takes a case study approach. The research reported here uses data collected from three cohorts of students enrolled at Florida State University. The data include information about each student's performance at high school, ethnicity, gender and first year GPA. Multilevel modelling has been applied to the data using the statistical software package MLwiN[1] to investigate the relationships between the variables, and in particular to determine which are the best indicators of academic success at university, whilst taking into account the effects of individual high schools. Issues relating to choice of predictive and university success measures, intervening variables, controlling for selection bias, data and measurement, and choice of research model will be discussed in the context of an American university.

U.S. secondary school indicators for success

Given the increase in the number of applications for admissions to colleges and universities for the limited number of seats in freshmen classes, students and universities in the U.S. must consider all available indicators for success in higher education. There are many ways a student can gain recognition to contribute towards their university application. The standard high school exam in the U.S. is the SAT (formerly known as the Scholastic Aptitude Test) although in some states an alternative, the

1. www.cmm.bristol.ac.uk/index.shtml

2. Concordance tables are published to find equivalences so that SAT scores can be used for the minority of students who take the ACT.

ACT (American College Testing), is more popular[2]. In this study we are considering students in Florida, where the majority take the SAT exam. Although standardised test scores have varying significance in the admission decisions of all students who qualify for admission at universities in the U.S, all potential U.S. university students must submit results of college entrance exams, either SAT or ACT, in order for an application to be considered complete in many universities. In addition to this, students can choose to take additional exams, such as those that are part of the Advanced Placement (AP), the International Baccalaureate (IB) or Cambridge's International A level programme (AICE)[3].

Advanced Placement has been a staple in U.S. education for over fifty years. Designed to promote excellence in secondary education, the programme desires to allow motivated students to work at their optimum capability. Nearly one million U.S. students now take at least one AP exam during their secondary careers. As Harvard, Yale and Princeton Universities were active participants in the study that led to the creation of AP, the acceptance of this credential is nearly universal among American universities.

In the late 1960s the International Baccalaureate was founded. While initially established as a single programme for internationally mobile students, the programme has flourished throughout the world, but nowhere greater than in the U.S. By 2005 over 1,000 secondary schools in North America offered the IB curriculum. The IB had to work diligently to have U.S. universities provide recognition similar to that provided to AP.

While Cambridge has been offering examinations for 150 years, it is relatively new in offering its curriculum in the U.S. The four year IGCSE/AS/A level curriculum and exams leading to an Advanced International Certificate of Education Diploma were introduced in Florida's Bay High School a little over fifteen years ago. Cambridge is experiencing the same curve of recognition as IB experienced in the 1970s and 1980s.

A tabulated comparison of secondary education in the UK and the US is shown as an appendix.

Explanations of terms used

For the benefit of readers who may not be familiar with the U.S. high school and university system we include here some explanations that may be helpful.

Cambridge Advanced International Certificate of Education Diploma: Cambridge awards a Cambridge AICE Diploma to students who have passed a prescribed number of subject examinations at the Advanced (A) level and/or the Advanced Subsidiary (AS) level. To qualify for a Cambridge AICE Diploma, students must pass at least one examination from each of three subject groups to include Mathematics and Sciences, Languages (both foreign and first), and Arts and Humanities. In the US, Cambridge International AS and A level examinations are sometimes referred to as 'Cambridge AICE' or 'AICE' examinations. Students passing AS and A level examinations may be awarded entry level or intermediary level university course credit by examination or advanced standing at US colleges and universities.

Advanced Placement: The AP programme is a curriculum in the US sponsored by the College Board[4] which offers standardised courses to high school students that are generally recognised to be equivalent to undergraduate courses in college. Participating colleges grant credit to students who obtained high enough scores on the exams to qualify. During their secondary studies a student may opt to take many AP courses, or as few as one. This curriculum is the most widely spread acceleration mechanism offered in the US and has been in place for over fifty years.

Credit hour: Each course that a student can enrol on is worth a certain number of credit hours. One credit hour is normally equivalent to 'one hour of classroom instruction and two hours of student work outside class over 15 weeks for a semester' so that a typical course is worth 3 hours, and this can vary from 1 to 5. Different institutions can vary how much credit is assigned to Cambridge AICE, AP or IB results.

Dual enrolment: Dual enrolment is normally concurrent enrolment where a high school student is taking a college course for both high school and college credit. This may be done by the student being released from his/her high school and taking the course on a college campus, or by the college approving the curriculum and allowing the student to remain on the high school campus and the college appointing the secondary school instructor as an adjunct faculty member at the college. Many students will earn a year of college credit in this manner, and some students will earn as much as two years of credit through dual enrolment. Many parents see dual enrolment as a money saving strategy to avoid high tuition costs at universities and state governments see this as a net saving since public school costs are lower than they would be at post secondary institutions.

High school GPA: High schools in the US determine how to calculate GPAs for purposes of generating a rank distribution. The system gives 4 points for a grade A, 3 points for a grade B and so on, and then takes the average, so that the final score is out of 4. (Given different weighting systems for advanced level courses, the GPA could exceed 4.) The lack of moderation in this process makes it more difficult to give standardised measures of high school performance, although there is evidence to suggest that HSGPA is nevertheless a good predictor (Betts and Morrell, 1999). One possibility is to sort students into categories based on their rank.

International Baccalaureate: The IB diploma programme is offered at over 3,000 schools in over 130 countries. The diploma programme is a two year programme and to receive an IB diploma a student must complete courses in social studies, mathematics, experimental sciences, their primary language and a second language. A sixth course must also be completed with a choice of an arts course, or a second course from the five disciplines mentioned above. In addition to the six courses, students must complete an extended essay, complete a course titled 'Theory of Knowledge' and complete a requirement of activity beyond the classroom. Three courses must be completed at the Higher Level while the other three can be taken at the Standard Level. College credit and placement may be earned, although the amount of credit and the score necessary to receive credit will vary by institution.

No Credit: Nearly all US high schools have what is commonly referred to as a 'college preparatory' curriculum. This curriculum is designed to

3. http://www.cie.org.uk/qualifications/academic/uppersec/aice
4. The College Board is a not-for-profit membership association in the US that was formed in 1900 as the College Entrance Examination Board (CEEB) www.collegeboard.com
5. http://www.universityworldnews.com/article.php?story=20100625183517482

prepare a student for successful study at the college level. If no credit is included that could mean that no acceleration mechanism such as Cambridge AICE, IB, AP or dual enrolment has been included in the course of study or the student took an AP/IB/AICE curriculum, but did not score sufficiently to receive credit.

SAT and ACT scores: Almost all students take either the SAT exam or the ACT exam, and some take both. The SAT was revised in March 2005. The revisions were made to enhance the test's alignment with current high school curricula and emphasise the skills needed for success in college (see Lawrence, Rigol, Van Essen, and Jackson, 2003, for a detailed explanation of the changes).

The SAT is composed of three exams:

- Critical reading (SAT-CR)
- Mathematics (SAT-M)
- Writing (SAT-W)
- Total (SAT-Tot)

The score scale range for each section is 200 to 800 and the score scale range for the total is 600 to 2400. The official SAT website[6] states that, for 2006, a total score of 1800 means the candidate scored better than 80.8% of test takers. Admittance into many highly regarded American colleges requires scores above 1800, although entry will also depend upon a student's academic transcript (record of academic achievement) and extracurricular activities.[7]

Florida State University: a case study

This study takes a case study approach using data from Florida State University. Denscombe (2003) describes the key characteristics of case study research: spotlight on one instance; in-depth study; focus on relationships and process; natural setting; and multiple sources and methods. (For detailed explanations and discussions of case study research, see Denscombe, 2003; Bell, 2005; Cohen, Manion and Morrison, 2007; and Sharp, 2009.)

In general, case studies can be used to: (a) provide a thick description of complex interactions to enhance understanding of a range of social phenomena, (b) corroborate theoretical suppositions, and (c) generate and contribute to theory (Eisenhardt, 2002; Yin, 2006). Therefore, when giving consideration to case study methodology, it is necessary to understand it as "both a process of inquiry about the case and the product of that inquiry" (Stake, 2008, p. 121).

Florida State University (FSU) is a publicly supported institution located in the state capital of Tallahassee. FSU is a comprehensive, national graduate research university with 40,255 students of whom 8,557 are graduate students. FSU is home to the National High Magnetic Field Laboratory and their arts programme – dance, film, music and theatre – is widely regarded within the U.S. Recently FSU added a College of Engineering and a College of Medicine. The university also has a College of Law.

6. www.satscores.us
7. Interpreting SAT Scores and ACT Scores. University Language Services. http://www.universitylanguage.com/guides/interpreting-sat-scores-and-act-scores/

Exploring the issues

In what follows we outline some of the issues relating to the implementation of a predictive validity study in the context of an American university.

Choice of predictive success measure

A challenge to all models interested in prediction is the choice of predictive success measure.

The College Board encourages universities to use SAT and high school grades when making admissions decisions. However, high school grades are not necessarily a good means of comparing students' experiences and achievements prior to university. This is because high school grades reflect the standards and quality of a particular school or schooling system. These standards differ according to school area or region (e.g. urban or rural) and even individual schools. Moreover, inter-school effects are not always reflected in high school grades (Burton and Ramist, 2001).

The primary purpose of the SAT is to measure a student's potential for academic success in college. In this context, a number of studies have been undertaken which attest to the predictive validity of the SAT. (For a useful summary relating to the predictive utility of SAT, ACT and high school GPA [HSGPA] as indicators of university success see Cohn, Balch and Bradley, 2004.)

Cohen, Manion and Morrison (2007) used SAT scores, HSGPA and high school class rank to determine how well these predict college GPA. Data were collected from 521 students enrolled on Principles of Economics at the University of South Carolina in 2000 and 2001. They examined the frequency distribution of key variables and regression analysis (no multilevel model), with students grouped according to gender and race. It was found that having a SAT score of over 1100 (out of a possible 1600) and a class rank of over 70 gave a predicted college GPA of around 3.0.

A large-scale national validity study of the revised SAT (incorporating an additional section in writing and minor changes in content to the verbal and mathematics sections) was undertaken by Kobrin, Patterson, Shaw, Mattern, and Barbuti (2008). Their studies were based on data from 150,000 students from 110 four-year colleges and universities across the US entering 110 four-year colleges and universities in the fall of 2006 and completing their first year of college in May/June 2007. The writing section was shown to be the single most predictive section of the test for all students. The analyses also found the writing section to be the most predictive across all minority groups. The studies also revealed that:

- SAT is an excellent predictor of how students perform in their first year at university;
- SAT is a stronger predictor than high school grades for all minority groups (African American, Hispanic, American Indian and Asian);
- the recently added writing section is the most predictive of the three SAT sections.

Culpepper and Davenport (2009) studied a sample of 32,103 first-year students who were enrolled in one of 30 colleges or universities in 1995. They compared the attainment of students from different racial/ethnic backgrounds, and found that an African-American student with the same HSGPA, SAT or ACT score as a white student was likely to have a lower college GPA. The possible differential prediction of SAT scores for university performance by race highlights the need to control for race in models involving SAT scores.

However, not all studies have produced evidence that the SAT

identifies the students most likely to succeed at university. Lenning (1975) carried out three studies to determine whether ACT was as good a predictor of college grades as SAT for highly selective institutions. Although only three such institutions were studied, they found that ACT scores can be at least as predictive, and likely more predictive, of college grades at highly selective institutions than SAT scores.

Noble and Sawyer (1987) considered the ACT scores and HSGPA for students enrolled at 233 institutions across 2812 courses in October 1985. They computed regression statistics for each course. They found that including HSGPA gave a stronger prediction of college GPA.

Noble (1991) conducted a study of 30 colleges, mainly located in central and southern U.S, with a higher than representative proportion of public colleges. It was found that ACT is a reasonable predictor of college success, and that including HSGPA improves the predictive validity.

A study by Betts and Morrell (1999) also indicated that HSGPA (as well as SAT scores) are significant predictors of university GPA.

Choice of university success measure

Another challenge to models interested in prediction is the choice of university success measure. For example, a number of different university performance measures could be used. These may include:

- average GPA for first year (or other years if available)
- number of courses passed
- number of courses excelled in
- GPA in certain courses, for example, science/mathematics versus humanities
- university enrolment status (as of the second fall after high school graduation)
- university retention, that is, re-enrolment in a second year at the same institution (Robbins, Allen, Casillas, Peterson, and Le, 2006)
- certain measures of engagement, for example, more propensity to participate in research at university or study abroad, more likely to participate in a student activity of some kind, etc.

However, the ultimate choice of performance measure would depend on data available and whether the data provide a comparable measure across courses included in the study.

The concept of tertiary level academic success used here is determined by the persistence of a student within the university with a specific GPA. The definition of university GPA employed is based on the accumulation of all previous semesters' work. In this study we are considering the GPA for students attending just one university. However, future studies will entail collecting data from a number of universities which may create different challenges. For example, it would appear that U.S. universities demonstrate some degree of latitude in determining how to calculate GPAs.

Choice of research design and hypotheses

In order for the research to be well-founded, we must ensure that:

- the analysis of statistical indices (e.g. correlations, regression coefficients) is technically sound and in particular that it:
  - addresses a set of testable hypotheses, derived from a sound theoretical approach, and
  - uses appropriate empirical methodologies and data for the purpose
- any inferences drawn from the analysis are justified and that erroneous inferences in the public domain (as may be drawn by third parties) are either avoided, or otherwise addressed and corrected as appropriate.

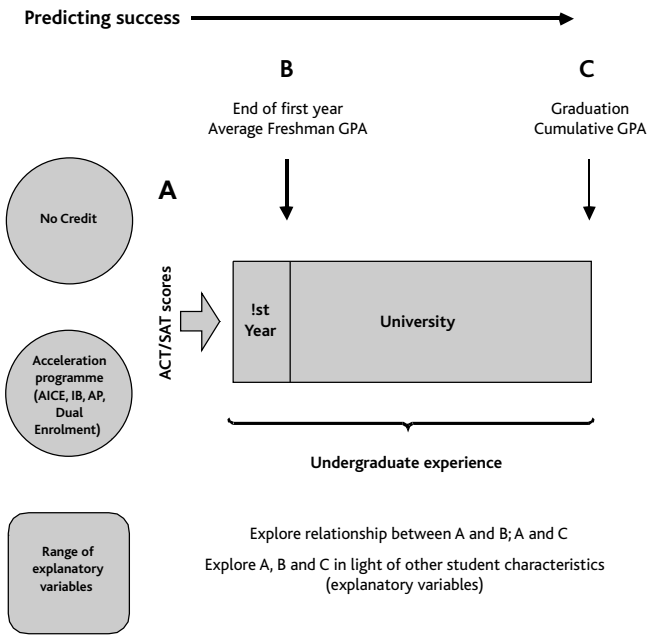
The principal hypothesis tested in this initial, exploratory study may be stated in the following way:

> *Students who follow the AICE, AP or IB programmes will achieve a significantly higher first year GPA than those with no credit, given the same SAT scores.*

The research designed to test this hypothesis may entail the formulation of several preliminary model specifications (each based on unit data where each student represents a single observation).

In order to estimate predictive validity it is necessary to determine the relationship between the success of students leaving high school following a particular programme of study and their success during, or at the end of, undergraduate study. Such a model is shown conceptually in Figure 1.

Figure 1: Predictive validity research design



A number of other models also have potential. For example, a test of predictive power using students who sit common examinations (i.e. a within-subjects design) – the hypothesis being that one assessment explains more variation in their university performance.

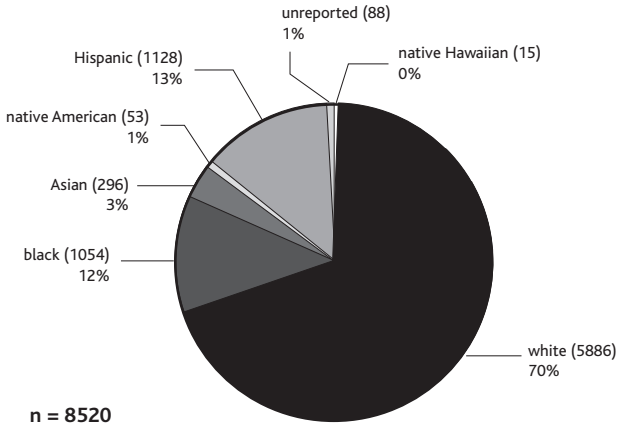
Choice of data and measurement model

The SAT score (total SAT score, SAT-Tot) has been used here as the choice of measure for high school performance. A point worthy of note is when students take the SAT. If students take the SAT late junior year or early senior year, then any additional acceleration programme may have an effect on their score.

To fit the multilevel models we used data based on records of over 8500 students who entered Florida State University during the academic years 2007/2008, 2008/2009 and 2009/2010.

Four datasets representing secondary educational programs were obtained from enrolment and admissions staff at the university. The largest data set (n = 6382) contained information on students with only

Figure 2: Pie chart to show the proportion of students of each race



the SAT (or ACT) score (hereafter referred to as having 'no credit'). The three other data sets contained information on students with Cambridge AICE credit (*n* = 144), with AP credit (*n* = 1188) and IB credit (*n* = 806). Figure 2 shows student data in terms of relative proportions by race.

Column headings for each of the four datasets include: FSU student number, year enrolled, race, gender, FSU GPA, high school GPA, SAT verbal, SAT math, SAT total, ACT (if applicable), high school attended, type of exam program followed (if applicable). The explanatory variables are set out in Table 1.

Table 1: Explanatory variables definition

Generic data requirements	
Variable	*Explanation*
FSU student number	Unique student identifier
Race	1 = white, 2 = black, 3 = Asian, 4 = native American, 5 = Hispanic, 6 = unreported, 7 = native Hawaiian/other Pacific islander
Gender	M = male, F = female
FSU GPA	Possible values from 0 to 4
High school GPA	Possible values from 0 to 4 (or in some cases more than 4)
Matriculation year	Year first enrolled at FSU
SAT verbal	SAT score for critical reading component
SAT math	SAT score for math component
SAT total	Total SAT score
ACT composite	ACT score
High school code	Local high school identifier
Type of credit	Exam program followed – Cambridge AICE, AP, IB or no credit
Credit hours	Number of hours credit gained on a college course

The four data sets were combined into an overall matrix. The structure of the data, which contain students from (i.e. 'nested within') a number of high schools, suggests the use of multilevel models. The multilevel software package MLwiN (Version 2.02 Rasbash *et al*., 2005) was therefore used.

Multilevel modelling is a way of finding a line of regression through different groups, nests or hierarchies of data (unlike standard multiple regression techniques which assume that the observations are independent, which is not the case here). Multilevel models recognise the existence of both hierarchical data and clustered data structures.

Multilevel modelling takes account of the context in which a variable exists. It is often used in sociological applications because individuals are affected by, or defined by, the groups they belong to. For example, patients receiving the same treatment for the same condition at different hospitals may experience different patient outcomes; students in different classes or in different schools may obtain different exam results (outcomes). A two-level model which controls for student outcomes within high schools would include residuals at both the student and school level. In effect, residual variance is separated out into an inter-school constituent (the variance of the school-level residuals) and an intra-school constituent (the variance of the student-level residuals). The school residuals ('school effects') represent unobserved high school characteristics that affect student outcomes, more particularly student performance. The unobserved variables lead to correlation between outcomes for students from the same school.

Recognising how groups of individuals can be nested can help build a more realistic picture, giving insight into where and how effects are happening, and this is what multilevel modelling aims to do (see Goldstein, 2011; or Bryman and Hardy, 2009, for a more detailed description of multilevel modelling).

Not using a multilevel model as a result of failing to recognise hierarchical structures makes it more likely that a significant difference is reported when in fact the difference is non-significant (i.e. a false positive or type 1 error): standard errors of regression coefficients will be underestimated, leading to an overstatement of statistical significance. Standard errors for the coefficients of higher-level predictor variables will be the most affected if the effect of grouping is ignored.

As the outcome variable (FSU GPA scores – first year examination marks) is continuous, the model fitted was:

$$y_{ij} = \beta_{0ij}x_0 + \beta_1 x_{ij}$$
$$\beta_{0ij} = \upsilon_{0j} + \varepsilon_{0ij}$$

where y_{ij} is the predicted outcome variable (FSU GPA score) for individual i in high school j, β_{0ij} is a constant, β_1 is the independent contribution of the predictor variable to the dependent variable, x_{ij} is a predictor variable, υ_{0j} is high school level residual error and ε_{0ij} is individual level residual error.

Multilevel models have been used in several predictive studies to take into account the hierarchical structure of educational assessment data. For example, Bell and Dexter (2000) used multilevel modelling to investigate the comparability of GCSE and IGCSE and suggested that a wide between-school variation can make results misleading. However, this is the first study to our knowledge that uses multilevel modelling to compare the predictive validity of different types of high school exam programmes in the US.

Initial findings

Figure 3 shows the total SAT scores and the FSU GPA for each student in the dataset according to the exam programme followed. It can be seen that there are a number of outliers at the FSU GPA level – students who perform well in their SAT score but who do not do so well in their first year of college. In every case where students exhibit a zero score for their

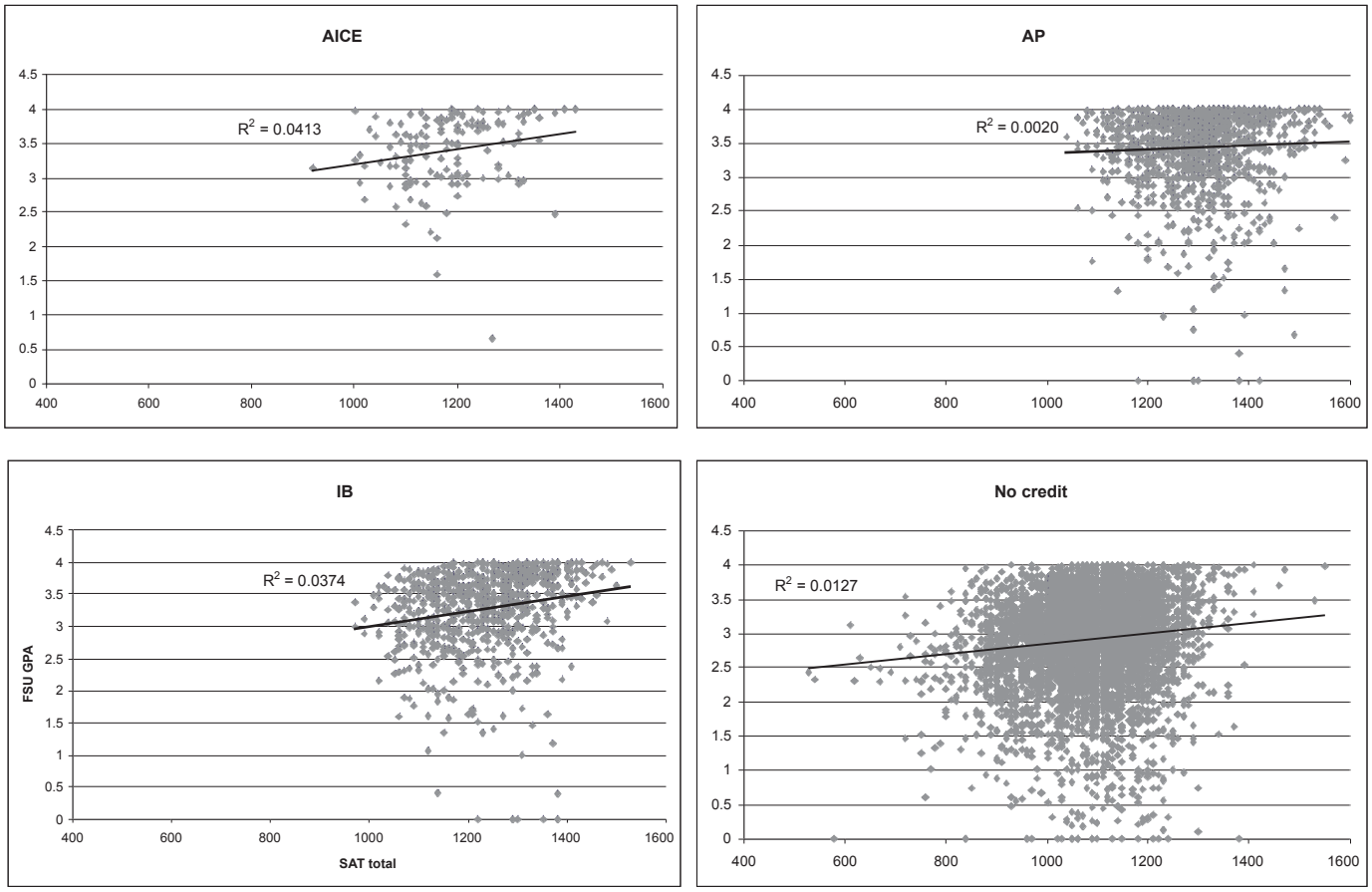


Figure 3: Scatter plots of the four datasets for each type of exam programme, showing SAT-Tot against FSU GPA and the line of regression and r^2 value

GPA it was noted that these were new students yet to receive a GPA. According to university admissions staff, any instances of low GPA scores are representative of underperforming students experiencing academic difficulties. It may be assumed, therefore, that these are special cases which a model could not reasonably predict. Consequently, any student with a GPA of less than 1.0 was excluded from the data set. It should also be noted that most of the student GPAs shown in Figure 2 fall within the range 2–4 (though this range is wider for 'no credit' students).

The SAT scores for students with no credit are considerably lower than those of the other three groups.

Using the refined dataset (excluding FSU GPA scores less than 1.0 and with the 488, or 5.7% of candidates missing SAT-Tot scores replaced with equivalent ACT) the model investigates the factors associated with the course of programme study (Table 2). Regression coefficients are statistically significant if they equal twice or more the value of the standard error (shown in brackets). Statistically significant effects are shown in bold type. It should be noted that school-level effects appeared to be much smaller than the individual-level effects: there is no statistical difference between schools.

Table 2: Effect of educational programme (given equivalent SAT scores) on FSU GPA

Base – No credit	*Regression Coefficient (Standard Error)*
AICE	0.351 (0.053)
AP	0.359 (0.023)
IB	0.222 (0.026)

Compared to students with no credit (and controlling for the effects of SAT scores, gender and race), having taken the AICE, AP or IB programmes were all associated with significantly higher first year GPAs.

- Students who took the AICE attained, on average, a GPA of 0.35 higher than those with no credit, given the same SAT score.
- Students who took the AP attained, on average, a GPA of 0.36 higher than those with no credit, given the same SAT score.
- Students who took the IB attained, on average, a GPA of 0.22 higher than those with no credit, given the same SAT score.

Discussion

The aim of this study has been to determine how well acceleration programmes in the U.S. prepare students for success at university. This general question can be extended: by using multilevel modelling, we can ask how well a given exam programme prepares a student who comes from a particular educational background. The study has explored the link between high school quality (in terms of programme followed) to first year university academic achievement using data supplied by Florida State University.

Consideration of the issues and exploratory analysis of the data collected so far has enabled us to test whether students who follow the AICE, AP or IB programmes achieve a significantly higher first year GPA than those with no credit, given the same SAT scores and controlling for the effects of race and gender. The results show that following an examination programme results in, on average, a better GPA than not following any extra credit.

Validity considerations

On the inclusiveness of validity, Bachman has argued that it is important to recognise that no one type of validity evidence by itself "is sufficient to demonstrate the validity of a particular interpretation or use of test scores" (1990, p.237). Validity is a multi-faceted concept requiring a range of types of evidence to support any claims for validity of scores on a test: "These are not alternatives but complementary aspects of an evidential basis for test interpretation" (Weir, 2005, p.13). However, for studies of this kind predictive validity work must take priority for tests designed for use in university selection if the tests are to be seen as fit-for-purpose.

According to Weir (2005), establishing predictive validity through correlating secondary school performance or standardised tests against later academic performance is impeded by practical and logistical difficulties. Such problems are particularly pronounced when implementing tracer studies and also when attempting to identify and control for a range of confounding intervening variables (See Banerjee, 2003, for a critique of approaches to establishing predictive validity.) Conceptually, therefore, any predominantly quantitative and *a posteriori* estimation of validity should be triangulated with qualitative data collected from, for example, individuals within one of the main stakeholder groups: the learners and their teachers. There is a requirement for any examination board to demonstrate and share how they are seeking to meet the demands of validity in their assessments and to make every systematic effort to ensure that their assessments achieve a positive influence or impact on general educational processes and on the individuals who are affected by the results. Predictive validity and impact studies are important contributions, therefore, to the validation process of any assessment.

Weiss defines impact – from the perspective of educational evaluation – as "the net effects of a programme (i.e. the gain in outcomes for program participants minus the gain for an equivalent group of non-participants)" (1998, p.331). Acknowledging the narrowness of this definition, Weiss broadens its scope by adding that "impact may also refer to program effects for the larger community … more generally it is a synonym for outcome". Investigating impact is regarded as being an essential aspect of determining the utility (or usefulness) of an educational assessment in terms of fulfilling its intended purpose, that is, its fitness for specific purposes (validity broadly interpreted) and contexts of use. Embedded within the concept of impact reside the notions of *processes* as well as *outcomes* (or products). Roy (1998) distinguishes between the two:

> *A study of the product is expected to indicate the pay-off value while a study of the process is expected to indicate the intrinsic values of the programme. Both are needed, however, to find the worth of the programme.* (1998, p.71)

As there are a number of variables that can weaken the reliability of the conclusions drawn from this study, it is intended that the findings from a series of US impact studies will be used to support any predictive validity estimates.

It is important to the interpretation of any predictive research, therefore, that impact data collection instruments and procedures (such as questionnaires and interview schedules) are used in order to understand the test impact better and to conduct effective surveys to monitor it (Hawkey, 2004). Currently data are being collected in order to ascertain stakeholder perceptions of Cambridge assessments in the US educational system. School lesson observations together with semi-structured interviews and focused discussion groups with both students and teachers have been conducted in an attempt to gather information on pedagogic practice, lesson content, learning/study approaches and perceived features of test validity and reliability. These data have been enlarged and enriched through the collection of views provided by Higher Education admissions and teaching staff on how examination results are used and how secondary educational study programmes provide an indication of tertiary level preparedness and success. It is hoped that by undertaking longitudinal research and eliciting participants' perspectives on their own behaviour, a number of recurrent patterns across data sets will emerge thereby revealing "multiple aspects of a single empirical reality" (Denzin, 1978). Such an approach will provide Cambridge with greater clarity regarding their own assessments in terms of "what goes on while a program is in progress" and "the end results of the program" (Weiss, 1998, pp.334–335). Impact research will enable a closer exploration of the relationship between the experience of students in the Cambridge curriculum and the level of preparation for college as well as the level of success at college.

Study limitations

The focus of the research has been a case study. Case studies include both a process of inquiry that is grounded in interpretations and a contribution to a product from that inquiry. Although a case study methodology is not without its criticism (being a bounded investigation which suggests that products are not readily generalizable), "compared to other methods, the strength of the case study method is its ability to examine, in-depth, a 'case' within its 'real-life' context" (Yin, 2006, p.111).

A case study approach uses a constructivist/interpretivist orientation toward data collection and analysis processes. A case study methodology recognises the need for:

- multiple perspectives (as evidence that contributes to case descriptions); and
- multiple methods (in order to isolate and scrutinise perspectives within case studies).

Its adoption, therefore, is justified as a mode of situated inquiry, favouring uniqueness over generalizability.

The size of the dataset was large – over 8 500 students. This means the reliability we can attach to the findings is increased. Even where the sub-sets were small – for example, of Cambridge AICE students there were 144 – they were still sufficiently large for the analyses to be carried out. There were some sub-sets that were small, for example native American and Hawaiian, which increases the risk of Type II errors. (This is the error of failing to observe a difference when in truth there is one – a false negative.)

A common challenge in studies of this type is controlling for selection bias. The choice of educational programme is not necessarily random. High schools have different characteristics and in mixed Cambridge /non-Cambridge high schools students may have a choice. Students also may choose a high school based on its use of programme. To control for such potential bias, it would be useful to have some control variable that is correlated with the choice of system but otherwise unrelated to the student's performance at university. Typically we would expect the choice of system and student performance to be quite related. It is not clear what determines the choice of acceleration mechanism. Is choice of educational programme influenced by type of high school, extrinsic and

intrinsic motivational aspects, institutional ethos, affective characteristics, parental status, socio-economic constraints? Why do some students choose not to avail themselves of an acceleration programme? Clearly information of this kind would enhance our understanding of future predictive validity findings.

Future work

Further multivariate modelling work will include investigation of other variables which might explain student performance. Apart from a programme of learning these could include other students' characteristics such as socio-economic status, university enrolment status and university retention rates.

Other measures could include class type (whether Cambridge students do better with certain types of classes) or if certain behavioural measures, such as engagement with research or study abroad, might be enhanced. Apart from the freshman year cumulative GPA measure of achievement, other university performance outcomes could be explored, for example, four-year cumulative GPA scores; freshman year attrition rates; and four-year graduation rates. Additionally, it would be informative to compare SAT critical reading and SAT mathematics scores as there is some evidence that one is a better predictor of college success than the other.

All of the variables used for the above analyses come from university admissions records. Student transcripts from the administrative archives of the university provide information about university career (type and number of exam passed, frequency of study, credit hours, etc.) and data relating to some characteristics of the high schools attended (type of school, final grades). However, a questionnaire given to students when they enter university would enable the collection of additional information on the students' characteristics such as reasons for choice of educational programme and familial socio-economic status.

A valuable, longitudinal exercise would be to track an entire cohort of Cambridge students from one particular high school through to final year of study. Questionnaire surveys together with interviews throughout the duration of an AICE course could be undertaken in order to determine extent of workload, attitudes to course/assessment and teachers'/students' perceptions of the course. This would be accompanied by follow-up interviews with students at university, the findings from which could be triangulated with GPA scores achieved at the end of the first year of undergraduate study and also at graduation.

Given the smaller numbers in the AICE, AP and IB groups, the case study nature of the research and the possible presence of unknown confounding variables between groups, it would be unwise to draw conclusions about the relative predictive strength of the three acceleration programmes. Further work will be required to collect more data from both Florida State University and other U.S. universities. Cambridge has already obtained considerably smaller datasets from the universities of Maryland, Virginia and Michigan and the process of data collection is expected to continue over time.

Acknowledgements

We would like to thank John Barnhill (Assistant Vice President for Enrolment, Florida State University) and Megan Benson (Director, Enrolment Management Operations Office of Admissions and Records, Florida State University) for providing us with the university data and for their assistance and advice during the project. Thanks also to Sherry Reach (Cambridge US Regional Manager) and Bill Kolb (Cambridge US Recognitions Consultant) for their invaluable contributions throughout the course of this work.

References

Alderson, J. C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Anastasi, A. (1988). *Psychological Testing*. 6th edition. New York: Macmillan.

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

Banerjee, J. V. (2003). *Interpreting and using proficiency test scores*. Unpublished PhD thesis, University of Lancaster.

Bell, J. (2005). *Doing your research project: A guide for first-time researchers in education, health and social science*. 4th edition. Maidenhead: Open University Press.

Bell, J. F. & Dexter, T. (2000). *Using multilevel models to assess the comparability of examinations*. Paper presented at the 5th International Conference on Social Science Methodology, October 2000.

Betts, J. R. & Morrell, D. (1999). The determinants of undergraduate grade point average. *Journal of Human Resources*, **34**, 2, 268–293.

Bryman, A. & Hardy, M. A. (2009). *Handbook of data analysis*. London: Sage Publications.

Burton, N. W. & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980*. College Board Research Report No. 2001–02, College Entrance Examination Board, New York.

Cohen, J. & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the behavioural sciences*. 2nd edition. Hillside, NJ: Lawrence Erlbaum Associates Publishers.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. 6th edition. Abingdon: Routledge.

Cohn, E., Cohn, S., Balch, D. C., & Bradley, J. (2004). Determinants of undergraduate GPAs: SAT scores, high school GPA and high school rank. *Economics of Education Review*, **23**, 277–286.

Culpepper, S. A. & Davenport, E. C. (2009). Assessing differential prediction of college grades by race/ethnicity with a multilevel model. *Journal of Educational Measurement*, **46**, 2, 220–242.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of Language Testing*. Studies in Language Testing 7. Cambridge: UCLES and Cambridge University Press.

Denscombe, M. (2003). *The good research guide for small-scale social research projects*. 2nd edition. Maidenhead: Open University Press.

Denzin, N. (1978). *Research Act: Theoretical Introduction to Sociological Methods*. New York: McGraw-Hill.

Eisenhardt, K. M. (2002). Building theories from case study research. In: A. M. Huberman & M. B. Miles (Eds.), *The qualitative researcher's companion*. 5–35. Thousand Oaks, CA: Sage.

Goldstein, H. (2011). *Multilevel statistical models*. 4th edition. Chichester, UK: Wiley.

Hawkey, R. (2004). An IELTS Impact Study: implementation and some early findings. *Research Notes*, Issue 15, February 2004. University of Cambridge ESOL Examinations.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT® for Predicting First-Year College Grade Point Average*. Research Report, No. 2008–5. New York: College Board.

Lawrence, I. M., Rigol, G. W., Van Essen, T., & Jackson, C.A. (2003). *A Historical Perspective on the Content of the SAT*. Research Report No. 2003–3. New York: The College Board.

Lenning, O. T. (1975). *Predictive validity of the ACT tests at selective colleges*. Report No. 69 [050269000]. Iowa City, IA: American College Testing.

Noble, J. P. (1991). *Predicting college grades from ACT assessment scores and high school course work and grade information*. Report No. 91–3 [50291930]. Iowa City, IA: American College Testing.

Noble, J. P. & Sawyer, R. (1987). *Predicting grades in specific college freshman courses from ACT test scores and self-reported high school grades*. Report No. 87–20 [050287200]. Iowa City, IA: American College Testing.

Rasbash, J., Browne, W. J., Healy, M., Cameron, B., & Charlton, C. (2005). MLwiN Version 2.02. Centre for Multilevel Modelling, University of Bristol.

Robbins, S., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology*, **98**, 598–616.

Roy, S. (1998). A general framework for evaluating educational programmes. In: V. McKay & C. Treffgarne (Eds.). *Evaluating Impact*. 69–74. London: Department for International Development.

Sharp, J. (2009). *Success with your education research project*. Exeter: Learning Matters.

Stake, R. E. (2008). Qualitative Case Studies. In: N. K. Denzin & Y. S. Lincoln (Eds.). *Strategies of qualitative inquiry*. 3rd edition, 1–43. Thousand Oaks, CA: Sage.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.

Weiss, C. (1998). *Evaluation*. New Jersey: Prentice Hall.

Yin, R. K. (2006). Case study methods. In: J. L. Green, G. Camilli, P. B. Emore, A. Skukauskaite, & E. Grace (Eds.). *Handbook of complementary methods in education research*. 111–122. Washington, DC: American Educational Research Association/Lawrence Erlbaum.

APPENDIX: Comparison of secondary education in the UK and the US

	UK				*USA*			
Age	*Type of Institution*	*Year*	*Main Examination*	*Comments*	*Type of Institution*	*Grade*	*Main subjects/ examination*	*Comments*
14–15	SCHOOL	10		First year of GCSE course	HIGH SCHOOL	9	5 core subjects plus electives	• Students gain a Diploma in G12 • Credits for core and elective studies • Minimum number of credits needed; in Florida 24 • Many G11/12 pupils on Advanced Placement (AP) or Dual Enrolment (DE) as partof the credits • SAT taken in G11 and again in G12 if not good enough
15–16	"	11	GCSE (6–11 subjects)	Vocational courses also possible	"	10	5 core subjects plus electives	
16–17	SIXTH FORM or COLLEGE	12	AS (4–5 subjects)	Entry based on good grades in 4/5+ GCSEs	"	11	5 core subjects plus electives	
17–18	"	13	A2 (3 subjects)	The 'best' three AS subjects	"	12	3 core subjects plus electives	
18–19	UNIVERSITY	FIRST	First Year	Entry based on AS/A2 grades or points equivalent	COLLEGE	FRESHMAN	LIBERAL STUDIES	• Entry based on High School grades converted into GPA plus SAT score (plus in Florida community service) • They apply before receiving their Diploma • Offer based on minimum GPA + SAT scores in G12 • c.20% of students go to college
19–20	"	SECOND		"	"	SOPHOMORE	ASSOCIATE DEGREE	
20–21	"	THIRD	BACHELOR DEGREE	"	"	JUNIOR		
21–22	"	ONE	POST GRADUATE	Entry based on good first degree	"	SENIOR	BACHELOR DEGREE	

Evaluating the CRAS Framework: Development and recommendations

Martin Johnson and Sanjana Mehta Research Division

Introduction

This article reviews conceptual issues surrounding comparisons of demand through a critical evaluation of the CRAS (Complexity-Resources-Abstractness and Strategy) framework (Pollitt, Hughes, Ahmed, Fisher-Hoch and Bramley, 1998). The article outlines the origins of the CRAS framework in the scale of cognitive demand (Edwards and Dall'Alba, 1981). The characteristics of the CRAS framework are then outlined, with attention being drawn to the assumptions that underlie these characteristic features. The article culminates in a set of recommendations and guidance that are relevant for potential users of the CRAS framework.

The development of the CRAS framework

The CRAS framework (Pollitt *et al*., 1998) is an adaptation of an earlier scale of cognitive demand (Edwards and Dall'Alba, 1981). The Edwards and Dall'Alba Scale of Cognitive Demand was developed to evaluate lower-secondary level science materials. The primary purpose of the scale was to assess the cognitive demands set within the objectives, the learning tasks, and the evaluation instruments or techniques available to educators and to allow them to evaluate the internal consistency of cognitive demands across these different components. The theoretical foundation of the tool development process was eclectic, drawing on a number of learning theories including Bloom; Bruner; de Bono; and Novak's (1977) interpretation of Piaget.

For Edwards and Dall'Alba the cognitive demand of a task is based on the interaction of four dimensions: complexity; openness; implicitness; and level of abstraction. Moreover, within each of these four dimensions, six levels of demand were defined. The original scale is shown in Figure 1.

Trialling of the scale showed that the tool was useful when teachers reviewed a broad range of educational materials, enabling them to determine the degree of correspondence between their intrinsic cognitive demands. Furthermore, this trialling suggested that the tool was perceived to be advantageous in a number of respects, for instance, its application could lead to:

> *…awareness of features that may otherwise be overlooked; a more accurate and objective reflection of the materials…. and, revelation of the extent to which student performance on the evaluation instruments accurately represents their mastery of what it was intended they learn.*
> (Edwards and Dall'Alba, 1981, p.164)

The Edwards and Dall'Alba scale of cognitive demands was a primary influence on the development of the CRAS scales, which were specifically constructed to examine the effects of structure on demands in GCSE and A level examination items. Pollitt *et al*. defined demands as:

Figure 1: The Scale of Cognitive Demand: Edwards and Dall'Alba 1981

Characteristic Elements of Groups on the Scale				
Dimensions of Cognitive Demand				
Group	*Complexity*	*Openness*	*Implicitness*	*Level of Abstraction*
1	Simple operations	No generation of new ideas	Data are readily available to the senses	Deals with concrete objects or data stored in the memory
2	Require a basic understanding	↕	Data to be operated on are given	Predominantly deals with concrete objects or issues
3	Understanding, application or low level analysis	Limited generation of new ideas	A large part of the data is given but requires generation of the final outcome	↕
4	** ↕	Generation of ideas from a given data base		Corresponds to concrete-abstract transition
5	Analysis and/ or synthesis	Generation of ideas which are original for the student	Data are not available in a readily usable form – must be transformed	Abstract
6	Evaluation	Highly generative	Require a view of the entity in question as part of a more extensive whole	Highly abstract

** The arrows indicate that the characteristic element is intermediate between two more distinct points on the continuum.

> *…requests that examiners make of candidates to perform certain tasks within a question.* (p.6)

According to this definition, demands depend on the question and are the same for all candidates. Pollitt *et al*. articulate the relationship between the concepts of demands and difficulty more directly in their work when compared with Edwards and Dall'Alba. Pollitt *et al*. point out that these judgements of demand are necessarily made in advance of any knowledge about students' performances on such tasks and stand in contrast to their concept of difficulty. For Pollitt *et al.*, difficulty is represented by an empirical measure of how successful a group of students are on an item. In contrast to demand, which has no statistical indicator, difficulty can be explored through statistical techniques such as

'facility value', which "is the mean mark on a question expressed as a proportion of the maximum mark available – the lower the facility value the more difficult the question" (Pollitt *et al*., 1998, pp.105–106).

Pollitt *et al*. (1998) assessed the validity of an examination by comparing the demands set by the examiners in examination items to their overall impression of the responses to those items using the same CRAS scales. This task was undertaken to distinguish between the predicted demands that the examiners had intended when designing the items and the demands that were reflected in student performance on those same items. In this way the presence of the intended demands could be validated through a reflection of actual performance on the item. Without this post-hoc validation the predicted demands would remain untested and lack any ability to support their wider application.

Another adaptation of the original scale led to the inclusion of an additional dimension called 'strategy' into the new framework. The inclusion of this additional scale was supported by an augmentation of the theoretical base of the original Edwards and Dall'Alba scale.

Another contrast between the original Edwards and Dall'Alba scale and CRAS related to the number of levels of demand and the precision of their definition. The original Edwards and Dall'Alba (1981) Scale of Cognitive Demands consisted of a set of dimensions that ranged across six levels of demand. However, in the CRAS scales, the number of levels was reduced to five. In addition, the levels were more loosely defined. In comparison to the inclusion of explicit descriptions for 20 of the 24 dimension levels in the original scale, the new scales contained descriptions for only levels two and four of each dimension; amounting to eight descriptors in total. Hughes *et al*. (1998) suggest that these amendments were necessary to increase the flexibility of the scales, to move it away from its original science-specific context, and to allow judges (examiners) in other subject areas to use their professional *judgement* to make their own subjective comparisons.

These revisions resulted in the development of the CRAS framework which includes the dimensions of: complexity; resources; abstractness; and strategy (Figure 2).

Further revisions of the CRAS scales were then carried out to develop subject-specific scales for judging demands in examination items in History, Geography and Chemistry. Although acknowledging limitations of the CRAS framework in relation to affective and psychomotor demands, these revisions allowed the authors to claim that:

> *The scales can be used to see if the demands of the (i) text books and teaching materials, (ii) national curriculum, (iii) lesson content, (iv) assessment tasks, and (v) marking criteria, are matched.*
> (Hughes *et al*., 1998, p.18)

The features and assumptions underlying the Scale of Cognitive Demands and CRAS

Both sets of cognitive demand scales have a number of similarities and differences in relation to each other. It is important to compare the underlying reasoning which contributes to these similarities and differences.

This article uses two terms to help elaborate this comparison. The superficial and more obvious characteristics of the scales are termed 'features'. The paper goes on to argue that these features are intrinsically linked to sets of 'assumptions' which underlie them. In other words, assumptions are the logical underpinnings of the scales and which help to shape their features.

Figure 2: The CRAS Framework of Demands: Hughes *et al*., 1998

	← *Level* →				
Dimension	*1*	*2*	*3*	*4*	*5*
Complexity The complexity of each component operation or idea and the links between them	←	• Simple operations (i.e. ideas/ steps) • No comprehension, except that required for natural language • No links between operations	← →	• Synthesis or evaluation of operations • Requires technical comprehension • Makes links between operations	→
Resources The use of data and information	←	• All and only the data/information needed is given	← →	• Student must generate the necessary data/ information	→
Abstractness The extent to which the student deals with ideas rather than concrete objects or phenomena	←	• Deals with concrete objects	← →	• Highly abstract	→
Strategy The extent to which the student devises (or selects) and maintains a strategy for tackling and answering the question	←	• Strategy is given • No need to monitor strategy • No selection of information required • No organisation required	← →	• Student needs to devise their own strategy • Student must monitor the application of their strategy • Must select content from a large, complex pool of information • Must organise how to communicate response	→

This section sets out the features and assumptions for both scales. Once key similarities and differences in these features and assumptions are stated there is a brief outline of the claims that are made by each of the respective authors for each set of scales. The shared features (SF), divergent features (DF), shared assumptions (SA), and the divergent assumptions (DA) are described and evaluated in this section.

Shared features (SF)

SF1: The scales are based on an eclectic combination of educational theories

SF2: The scales are used to determine cognitive demand

SF1: The scales are based on an eclectic combination of educational theories

The original scale draws from a range of cognitive and learning theories: because CRAS is based on these original scales, it obviously draws on the same theories. At the same time, the authors of CRAS supplement the original theoretical foundations with more recent work in order to make the scales more applicable to their particular context (examination materials). It is possible that this process of theory building has some problematic elements.

The development of the original Edwards and Dall'Alba scale was based on selection, interpretation and amalgamation of specific theories. The authors justified this interpretative process by arguing that a single theory cannot be all encompassing; they needed to integrate a range of ideas. Although their justification appears reasonable, the exact process of selecting and combining elements from different theories is not entirely clear and raises an important question: can established theories based on their particular central tenets be aggregated in a single tool?

Research related to combining two or more theories into a single theory or conceptual framework is becoming relatively common. In the absence of all encompassing theories, researchers are increasingly identifying the need to construct broader frameworks by combining theories to study complex realities (Radford, 2008; Wedege, 2009; Strauss, 1986). It is suggested that the integration of theories should result in more holistic answers to certain research questions (Tsamir and Tirosh, 2008). Whilst it is accepted that theories originate in specific contexts and provide particular explanations for phenomena, it is also suggested that elements within different theories could complement each other to arrive at a feasible amalgamation (Strauss, 1986). However, it is very important to define the limits of this combination process in order to ensure that the revised theory remains meaningful and relevant.

The process of combining theories needs to be made transparent. More importantly, it also suggests that a researcher will have to carry out an evaluation of each theory that is being considered for integration in a larger framework to determine its goodness-of-fit in that broader framework.

Since CRAS is based on the theoretical framework of the Scale of Cognitive Demands (Edwards and Dall'Alba, 1981) which combined concepts and principles related to learning and cognition from a number of theories, it carries with it some of the ambiguities related to the original development. Whilst Edwards and Dall'Alba (1981) listed the sources from which each of their four demand dimensions were adopted or adapted, the rationale for this selection was not articulated in detail. In the absence of these details the theoretical conceptualisation of CRAS does not lend itself to a critique of the rationale for choosing between the different, and potentially competing theories that were, and that could have been included in the framework. It can only be concluded that combining concepts from different theories is possible, however, the appropriateness of the theoretical framework on which CRAS is established cannot be fully explored.

SF2: The scales are used to determine cognitive demand

Both the Edwards and Dall'Alba and the Pollitt *et al*. scales were created to assess the cognitive demands that are placed on students when engaging with particular tasks. Whilst Edwards and Dall'Alba tie their scale to the scientific learning domain, they suggest that scale application can be used with a diversity of source documents, for example, "The tool is used to determine the cognitive demand levels of the objectives, learning tasks, and evaluation, and to allow a comparison between these" (1981, p.160). On the other hand, Pollitt *et al*. suggest that their adaptation has less learning domain specificity but that it has a tighter focus on specific source documents, for example, for use with assessment items.

Both scales are based to some extent on the taxonomy of learning objectives developed by Bloom (1956). This taxonomy classified learning objectives into three domains, affective, psychomotor, and cognitive. It is notable that both the Edwards and Dall'Alba and CRAS scales focus exclusively on cognitive demands and choose not to engage with either affective, or psychomotor demands.

Divergent features (DF)

DF1: Scale length and level definition
DF2: Attending to constructs

DF1: Scale length and level definition

A contrast between the original Edwards and Dall'Alba scale and CRAS relates to the number of levels of demand and the precision of their definition. The original Edwards and Dall'Alba (1981) Scale of Cognitive Demands consisted of a set of dimensions that ranged across six levels of demand. However, in the revised Hughes *et al*. (1998) framework, the number of levels was reduced to five. In addition, the levels were more loosely defined in the new adaptation. In comparison to the inclusion of explicit descriptions for 20 of the 24 dimension levels in the original scale, the new framework contained descriptions for only levels two and four of each dimension; amounting to eight descriptors in total. Hughes *et al*. (1998) suggest that these amendments were necessary to increase the flexibility of the framework, to move it away from its original science specific context, and to allow judges (examiners) in other subject areas to use their professional judgement to make their own subjective comparisons.

DF2: Attending to constructs

It appears that the relationship of the two demand frameworks to the concept of construct validity differs slightly. In the development work related to the original Edwards and Dall'Alba (1981) scale there is explicit reference to the way that the content, and perhaps by association the constructs, of the science materials were attended to (1981, p.162). In the CRAS development work this link between demands and content/constructs is less clearly articulated.

Whilst the CRAS framework does not explicitly refer to the concept of construct validity in its dimensions it appears that the concept is implicit within the CRAS framework. Construct validity is a concept that test developers and evaluators need to consider. In the CRAS framework the link between demands and content/constructs appears to be more implicit than explicit. Reviewing an item using CRAS involves an analysis of demands in relation to those intended by the item developer. Any discrepancy between the intended and observed demands would indicate that there might be some potential for construct irrelevant variance which would threaten the validity of the item.

Shared assumptions (SA)

SA1: The interaction of multiple demand factors leads to the overall level of demand
SA2: The scales lead to a descriptive, qualitative account of cognitive demand
SA3: The scales enable evaluation of the internal consistency across the different demands
SA4: The scales can be used in conjunction with performance indicators to give insight into the relationship between demands and difficulty

SA1: The interaction of multiple demand factors leads to the 'overall' level of demand

Although both scales include slightly differing sets of dimensions, both conceptualise 'overall' demand in the same way. In line with Edwards and Dall'Alba's (1981) model, Pollitt *et al*. (1998) suggest that the demand dimensions within their CRAS model interact differently with particular features of an examination item. Since overall demand is based on the interdependence of the individual dimensions, changing one aspect of demand in an item might also alter the demands for other dimensions.

SA2: The scales lead to a descriptive, qualitative account of cognitive demand

Both sets of scales facilitate judgements about the demands of tasks which are essentially qualitative or descriptive in nature. Whilst this assumption is somewhat opaque in the work of Edwards and Dall'Alba, e.g. "[application could lead to]…awareness of features that may otherwise be overlooked; a more accurate and objective reflection of the materials" (1981, p.164), this perspective is more transparent in the development of CRAS: "The scales provide a language for examiners to articulate and share discussion, thus building an awareness of those demands…" (1998, p.18). An important implication of this shared assumption is that both scales aim to build a rich description of the demands inherent to a task.

It is important to highlight the point that the accounts generated through these demand frameworks remain at a general level. They do not offer insight into the variability between situations that might have influenced why there could be a difference between what an assessment item intended to do and how a student performed on it. Through triangulation of the projected demands inherent to assessment items, a curriculum, and a mark scheme, the two demand frameworks seek to present a general picture of demands. This analysis remains at the macro-system level and lacks a particular focus on the individual circumstances which might influence student performance. In other words, micro-level variances at teacher and class level within different schools are not a conceptual consideration of the CRAS or the Edwards and Dall'Alba scales. Users of these scales therefore need to bear these limitations in mind if they are interested in gaining such particular insights.

SA3: The scales enable evaluation of internal consistency across the different demands

The Scale of Cognitive Demands is based on the claim that it can be used to identify and compare cognitive demands across related educational components: objectives, learning tasks, and evaluation (Edwards and Dall'Alba, 1981). Similarly, the authors of CRAS claim that analysis of demands using CRAS across several components (text books and teaching materials; national curriculum; lesson content; assessment tasks; marking criteria) can be carried out to determine the degree of match (Hughes *et al.*, 1998). However, the authors of CRAS do not provide any further details on what may be the ideal level of correspondence between demands across these different components.

SA4: The scales can be used in conjunction with performance indicators to give insight into the relationship between demands and difficulty

Implicit to both sets of scales is a relationship between the demands of a task and its level of difficulty. Although this relationship is not considered to be direct, the use of the scales allows insight into the interplay between these two factors. Again, whilst Edwards and Dall'Alba are more vague than Pollitt *et al*. about the concept of difficulty in their work, it can be inferred that they do allude to the relationship between demands and difficulty, for example, "[application of the scales could lead to]…revelation of the extent to which student performance on the evaluation instruments accurately represents their mastery of what it was intended they learn" (Edwards and Dall'Alba, 1981, p.164).

Pollitt *et al*. (1998) conceptualise this relationship in greater depth through discussion of the use of structure in examination items. The term structure can be used to describe item features such as the layout and the number of steps of operations required. Pollitt *et al*. (2007) explain that structure is widely used by examiners to influence the demands of items, and by considering judgements about the demands in such items it is possible to investigate whether these structural features also have effects on any empirical measures of difficulty experienced by students when attempting such items.

Divergent assumptions (DA)

DA1: Item types that the scales can deal with
DA2: The breadth of contexts for scale use
DA3: The capacity of language to describe judgements
DA4: The relative importance of reliability or validity
DA5: The nature of the judgements supported by the scale
DA6: Combining scale judgements
DA7: The role of the scale user

DA1: Item types that the scales can deal with

The Edwards and Dall'Alba scale was designed for use with evaluation items that had objective or multiple choice characteristics. On the other hand, the CRAS framework was developed to be used with a more diverse set of materials. Hughes *et al*. highlight that the CRAS framework was developed to deal with examinations that incorporated a mixture of both structured and essay items (1998, p.18).

DA2: The breadth of contexts for scale use

The Edwards and Dall'Alba (1981) scale was specifically designed to deal with demands in the context of science materials. The CRAS framework was developed to be able to generalise across a variety of subject discipline levels. Hughes *et al*. state that the CRAS development process purposively involved three subjects (History, Chemistry and Geography) so that content coverage spanned "most of the disciplines (mathematical, literary, and physical and social scientific)" (1998, p.18).

DA3: The capacity of language to describe judgements

The Edwards and Dall'Alba scale includes clearly articulated statements along almost all of the points of the rating scales for each cognitive dimension. This implies that the authors believe that language has a capacity to adequately describe qualities of phenomena which can then facilitate judgements to be made against them. This use of rigidly defined criteria contrasts with the approach taken by Pollitt *et al*. for the development of CRAS. The CRAS framework opted to use only two defined scale points for each dimension. This difference in approach reflects Pollitt's concern that trying to use language to encourage absolute judgement making would be useless, since "language, like judgement, is inherently comparative and only approximately quantitative, and the problems of trying to pin down relative meanings with words are well known" (2007, p.189).

DA4: The relative importance of reliability or validity

The Edwards and Dall'Alba tool includes a highly defined cognitive demand scale, which implies that there is a great emphasis placed on how to support the reliable use of the scale. In light of this, a significant portion of the 1981 Edwards and Dall'Alba paper, describing the process of scale development, deals with the issues of establishing inter-rater reliability for use of the scale. Implicit in this process is the sense that attaining high levels of reliable scale use is predicated on good levels of scale user understanding of the scale descriptors. In this way, high reliability is indicative of high validity.

Pollitt *et al*., on the other hand, base their CRAS model on a set of "less stringently defined" cognitive demand scales at levels 2 and 4 of each of the dimensions (cited in: Hughes *et al*., 1998, p.5). The use of fewer descriptors in the CRAS model allows for the inclusion of elements that are relevant to scale users, thereby potentially enhancing the validity of the scale. At the same time, the existence of fewer descriptors heightens the importance of those remaining 'anchor' descriptors since these are needed to align the relative scales of different users into a common framework, since such a scale will always be "implicitly normed relative to the context in which it is being used" (2007, p.189).

Whilst Edwards and Dall'Alba largely avoid the problem of user interpretative variance with regard to the scale descriptors through clear articulation of each descriptor, the CRAS framework is less prescriptive in terms of the standardisation of user interpretation. This lack of prescription is important to highlight since any differences in scale ratings between two judges on CRAS should reflect 'real' differences in the stimuli being compared. An issue arises if inadequate understandings of scale points exist across judges since any variant outcomes might be indicative of differences in the stimuli being judged and/or differences between individual scale users' interpretations of the scales. The potential existence of these two sources of variance require different analytical approaches for scale interpretation than if only one source of variance was being observed (e.g. Cox, 1980, p.408).

DA5: The nature of the judgements supported by the scale

Because the Edwards and Dall'Alba tool comprises sets of clearly articulated statements at different levels of the scale dimensions there might be an inference made that this well-defined scale can support the making of absolute judgements of demand. This contrasts with the loosely defined CRAS scales, which reinforces the concept that individuals' judgements of demand are essentially relative in nature, that is, relative to other defined points on the scale.

DA6: Combining scale judgements

Again, the implied notion that the Edwards and Dall'Alba tool could help to capture 'absolute' judgements of demand has consequences on the potential combination of such judgement outcomes. Since there is an emphasis on the reliability of scaled judgements in the Edwards and Dall'Alba tool there is a suggestion that these judgements possess some mathematical or statistical characteristics. A consequence of this is that individuals' judgements might legitimately be combined in a quantitative fashion to give an overall level of cognitive demand.

This perspective contrasts very clearly with the Pollitt *et al*. view. Reinforcing the point that the dimensions of demand do not possess a quantitative structure Pollitt *et al*. state "despite the use of scales and the collection of numerical ratings the method is still fundamentally a qualitative methodology" (2007, p.192). The practical consequence of this is that "the results of a demand analysis will be to show that different exams make different demands…and it may be possible to say which demands each one requires most of, but it will usually not be possible to aggregate these validly to say that one is more demanding than the other" (2007, p.192).

DA7: The role of the scale user

The structure of relatively well-defined dimension scales in the Edwards and Dall'Alba tool supports its use across other cases, although only in relation to materials from within the context of Science for which it was developed. This contrasts with CRAS which contains loosely defined scales which are intended for use across different subject domains. This difference in structure and intended context means that the role of the scale user is somewhat different. For Edwards and Dall'Alba the well-articulated dimension scales and the clear context expectation constrains the user to ensure that the tool is applied appropriately. In relation to CRAS, the emphasis is on the tool user to establish whether their particular context is suitable for the application of the CRAS scales, and for the consequent modification of those scales.

Conclusion: recommendations and guidance for CRAS use

The identification of the divergent assumptions between the Scale of Cognitive Demands and CRAS is important as these help to explain the different features of the two scales of demands. Through its validation process the expectation of the Edwards and Dall'Alba scale developers is that it should be used as a tool in a very particular way and with little space for the scale user adaption. This contrasts with the CRAS framework since there is more emphasis on the users to adapt the scale for use in their own particular contexts, as long as they adhere to a number of key assumptions. In this way the CRAS scales operate more as a framework than a tool, with the framework resting on two key assumptions: first, that the four CRAS dimensions are used, and secondly, that the ability of judges to make relative judgements is supported by the scales.

This review of the assumptions and features of the CRAS framework leads to a number of recommendations and guidance notes for potential users of the framework. The links to these features and assumptions are referenced in parentheses.

1. The CRAS framework provides a common language to support teachers', examiners' and syllabus developers' conceptualisation and description of demands. The information elicited through the use of the CRAS framework, and the insights gathered, might be particularly important when working in a context where there is a lack of other evidence to draw on, for example, at the beginning of the development of a new assessment (SA2; DA3; DA4).
2. The CRAS framework is essentially qualitative in nature and can be used to profile the nature of cognitive demands for individual users. The rating for each dimension in one stimulus (e.g. an examination item) by an individual user can be used as a basis for comparison across other stimuli by the same individual. This comparison is meaningful because the user is making ratings according to the same underlying reference scale. It is not possible for an individual user to combine the ratings of each dimension to reach an overall 'level of demand'. This overall score is not meaningful as a basis for

comparing different stimuli because the interplay between the different dimensions might have compensatory qualities. By combining the ratings of different dimensions to arrive at a total score the user compromises the qualitative power of the CRAS framework, which aims to demonstrate that each stimulus has different demands and seeks to give the user a language to explicate the nature of those demands (SF2; SA2; DA3; DA6).

3. CRAS recognises the concept that comparisons are based on relative rather than absolute judgements. Moreover, reflecting the complexities of judgement-making processes, the valid and reliable use of the framework relies on there being unidimensional reference scales for each CRAS dimension. Ensuring that this unidimensionality is maintained is perhaps easiest when there is a single scale user, the assumption being that the user will assign meanings to the scale points in a consistent way when rating different stimuli. Whilst the use of a single rater might maximise the reliability of scale application, it might not satisfy the condition for the scales to generate generalisable outcomes. Where multiple judges are involved in making these judgements there needs to be adequate standardisation so that judges' scale use is underpinned by common understandings of anchor criteria. These ratings might then be collectively analysed or subject to numerical treatment, but these treatments need to be meaningfully related to the nature of the data (DA4; DA6; SA2).
4. CRAS may be used in conjunction with other measures (e.g. facility values) to assess the level of difficulty. CRAS can give an insight into the demands that might relate to final difficulty outcomes, but this relationship remains tentative.

   This uncertainty remains for a number of reasons:

   - It is not necessarily the case that there is a direct 1:1 relationship between the CRAS dimensions of demand and difficulty.
   - Initial estimates of demands might also fail to relate well to actual difficulty measures because the concepts identified in items are not recognised in the connected mark scheme. In such cases the identification of such internal inconsistency would be valuable insight.
   - There might be disagreement between the intended/anticipated demands of an item as perceived by a subject expert and those actually experienced by the test taker. This might be due to a number of reasons: there might be factors unknown to the expert, such as teaching effects, that might have influenced the test taker; there might be misapplication of anchor descriptors in the CRAS exercise; and there might be misjudgement on the part of the expert.

   As a result of some of these factors the outcomes generated through a CRAS analysis will tend to be at the level of offering tentative insight into difficulty outcomes (SA3; SA4; DA2).
5. The CRAS framework relies on the users being able to relate their subject-specialist knowledge to the underlying features of the CRAS dimension scales. A precursor to applying the CRAS framework is the mapping of the dimensions to the area of study. This mapping process not only allows the users to demonstrate that the framework is fit for the context of the study, but it also allows adequate anchors on the dimension scales to be developed. This anchoring process is crucial for the scales to be used correctly. Subject-specialist knowledge level is also a crucial factor as this gives validity to the comparisons being made. If a CRAS user has knowledge that is unevenly balanced across two areas of study it will lead to invalid comparisons being made (DA7).
6. CRAS allows descriptions of cognitive demands to be made across a variety of subjects and qualifications. The potential range of application therefore is quite broad. As stated earlier, the relationship between CRAS and the area of study needs to be mapped. Once this mapping is complete CRAS can help to investigate whether the demands that were intended in an item are actually evident (SF2; SA2; SA3; DA7).
7. The rationale for using the CRAS framework is to investigate whether there is internal consistency between different elements of learning and assessment materials. In order to maintain conceptual clarity it would not be recommended that additional measures of cognitive demands be used in addition to CRAS. If a mapping exercise demonstrates that CRAS needs to be extended to include additional dimensions to deal with a context, this process is preferable to using additional sets of measures or alternative cognitive frameworks. By having a singular framework it is easier to compare measures across different elements to investigate internal consistency (DA7).
8. The original intentions of the CRAS framework were to give insight into the dimensions that contribute to item demand, with comparisons then being possible between different items according to their profile of demands. The CRAS framework is less clear about how these individual item characteristics interact when considered at question paper level, and how demands at an overall level might be conceptualised. What appears clear is that the concept of demands at an overall level would necessitate consideration of all of the items that comprise a question paper, and this would mean that selectively sampling items would be invalid.

   The original CRAS scales were used to rate the demands in single items or in item parts. Shifting away from this use might be considered problematic. In their original work Edwards and Dall'Alba make it clear that the cognitive demand of a task is governed by the interaction of different dimensions of demand:

   The level of cognitive demand of a task is determined by the interaction of all of its dimensions. (Edwards and Dall'Alba, 1981, p.159)

   One problem that flows from this is whether it is meaningful to combine sets of qualitative judgements into a 'CRAS score' for a whole paper. If a holistic profile for whole papers is generated by combining the demand scores for each component item, it is possible that the interplay of these item demands is overlooked. In other words, the interplay of individual demands within a question paper makes it problematic to try to combine all the multiple demands and relative compensations into a meaningful outcome which can be used as a point of comparison. For example, placing more or less demanding items at the beginning of an assessment can have an important impact on overall assessment demand; and this potential source of construct irrelevant variance is not captured by a simple aggregation of item demands to construct a measure of demands at a holistic paper level. Whilst it might be argued that CRAS can be used to compare singular items very well, the use of CRAS for multiple items leads to a superficial overview which gives little insight into how to resolve the multiple relationships between such items.

The CRAS dimensions might be used to give a language that can be used to glean an overall impression of the demands of a question paper, but this comparison will be somewhat superficial. Such an analysis will fail to elicit the particularities of the demands and their interrelationships that the framework was initially developed to capture (DA6).

References

Bloom, B. S. (Ed.) (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Susan Fauer Company, Inc.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, **17**, 4, 407–422.

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

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

Novak, J. D. (1977). *A Theory of Education*. London: Cornell University Press.

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

Pollitt, A., Hughes, S., Ahmed, A., Fisher-Hoch, H. & Bramley, T. (1998). *The effects of structure on the demands in GCSE and A level questions*. Report to Qualifications and Curriculum Authority. University of Cambridge Local Examinations Syndicate.

Radford, L. (2008). Connecting theories in mathematics education: challenges and possibilities. *ZDM Mathematics Education*, **40**, 317–327.

Strauss, S. (1986). Three sources of differences between educational and developmental psychology: resolution through educational developmental psychology. *Instructional Science*, **15**, 275–286.

Tsamir, P. & Tirosh, D. (2008). Combining theories in research in mathematics teacher education. *ZDM Mathematics Education*, **40**, 861–872.

Wedege, T. (2009). *Combining and coordinating theoretical perspectives in mathematics education research*. Proceedings of CERME 6, January 28th-February 1st 2009, Lyon France.

Developing a research tool for comparing qualifications

Jackie Greatorex, Sanjana Mehta, Nicky Rushton, Rebecca Hopkin and Hannah Shiell Research Division

Abstract

There are thousands of diverse qualifications in the UK. Comparability studies about qualification standards generally use the following as comparators:

- Quality of candidates' performance
- Demand

For new and vocational qualifications, samples of candidates' performance and assessment tasks (e.g. examination questions) can be small or unrepresentative and thereby inappropriate for research purposes. Consequently, researchers employ other comparators including *specification features*, e.g. depth of knowledge. The article details the process of devising a research instrument to compare the features of cognate units from diverse qualifications and subjects. Such an instrument is atypical but valuable for comparability studies.

As part of a wider project about comparing different types of qualifications Kelly's repertory grid interviews elicited knowledge from twelve experts. They represented three subjects and composite, general, vocational and vocationally related qualifications. A secondary thematic analysis of the data was completed. The result was a series of features:

- Learning
- Knowledge
- Summative assessment task
- Qualification system

Each feature had several sub-features. Both features and sub-features served to categorise the interview data. An instrument was derived from the features and sub-features, as well as the researchers' experience of qualifications. The instrument was refined through consultation with colleagues. The instrument in its final form consisted of a series of items relating to possible features of the different specifications. Respondents to the instrument were required to tick a box to indicate that the item applied to the given specification. See Appendix 1 for the full instrument.

A pilot of the instrument indicated that salient features vary somewhat between units. Therefore, as hoped, the research instrument highlighted the similarities and differences between units. This is the case for units of the same type and different types. However, there are no established conventions about how to analyse data. Therefore the instrument is suitable for use in future comparability studies about features, as long as the analysis of results is agreed from the outset. Future research might compare qualifications with data collected using the instrument.

Introduction

The aim of this article is to report the development of a research instrument. This is part of an ongoing project about methods of comparing specifications in a diverse qualifications system. For more details see Novaković and Greatorex (2011).

The instrument in its final form consisted of a series of items relating to possible features of the different specifications. Respondents to the instrument were required to tick a box to indicate that the item applied to the given specification. See Appendix 1 for the full instrument. The

research instrument is for comparing the specification features of cognate units from different types of qualifications. Results from the specification features instrument would highlight the similarities and differences between different specifications. These results might:

- help qualification users to make informed choices between specifications
- set the context for comparisons of what is more and less demanding in different qualifications
- be useful in the revision of specifications.

Concepts and terminology

It is important to consider some central concepts and terminology before explaining the process of developing the instrument.

The *specification (syllabus)* is a description of a qualification. Usually it contains the content (knowledge, skills and competencies), assessment arrangements, performance requirements, guided learning hours, suggested teaching arrangements and so on. A specification is the basis of a course intended to end in an award or certificate[1].

Specification features:

- are important characteristics of a qualification
- are deliberately built into qualifications
- might be explicitly stated in the specification
- might be part of the course intended by the specification
- apply to typical learners, rather than the most/least able learners or learners to whom special considerations apply.

For the remainder of the article, the specification features will be referred to as 'features'. Examples of features are breadth of knowledge and concrete knowledge.

At this stage a definition of features is given without a comprehensive list of illustrative examples of features. The research outlined below was conducted to develop such a list of features which will be the backbone of the research instrument in development.

Context: Qualifications system

The qualification system in England, Northern Ireland and Wales includes several types of qualifications. These include:

- General qualifications (GQs), which are usually academic qualifications. They incorporate the General Certificate of Secondary Education (GCSE) taken by most 16 year olds in England just before the end of compulsory schooling.
- Vocational qualifications (VQs), which are typically designed to recognise learners' competence in the workplace. National Vocational Qualifications (NVQs) are an example of VQs.
- Vocationally-related qualifications (VRQs), which tend to focus on an occupational sector and enhance learners' knowledge and prepare their readiness for employment.
- Principal Learning (PL), which are qualifications, but are also a part of Diplomas[2] along with units from other qualifications such as GCSEs, A levels, NVQs, Functional Skills and so on.

Since 1997 the National Qualifications Framework included all qualifications in England. Each qualification is assigned a level from entry level to level 8. Level 2 is the level prior to the end of compulsory schooling, level 6 qualifications include undergraduate degrees and level 8 qualifications include PhDs. More recently some qualifications were transferred to the Qualifications and Credit Framework (QCF). The QCF aims to show how the different types of qualifications inter relate and allow credit to be transferred between qualifications. It is a credit accumulation and transfer system. The QCF retains the nine qualification levels used in the NQF[3].

There are 139 awarding bodies and over 11,000 different qualifications. In some situations there is more than one qualification in a subject at a particular level that might serve as part of a pathway to further study or a job. For instance, there are 230 level 2 'art' qualifications; these include general, vocational and vocationally-related qualifications associated with 21 different awarding bodies. (For further details about the source of these figures see Appendix 2.)

In this qualification system, centres (schools and colleges) and learners choose between the available qualifications at a particular level. Additionally, admissions staff and employers decide which qualifications they will accept as indicating competence in a vocation or readiness for further study. Therefore, comparability studies which systematically map the similarities and differences might be useful (see Introduction).

There are some instruments which contribute to providing systematic information about features, see for example, QCA (2007a and b). No instrument has been developed (in the UK in the past decade) to compare features of cognate units from different types of qualifications and be suitable for re-use in various subjects. Therefore, these became the goals for the features instrument.

Research strategy

In summary, the three-stage strategy for developing the instrument was:

Stage 1: Identify features by conducting a secondary analysis of data from Kelly's repertory grid interviews with expert subject assessors. It was important to interview expert subject assessors about the specifications to gain their insights about the intentions of the specification as well as their constructs which take a subject assessment community perspective of the specifications.

A document analysis of the specifications by researchers who do not have the subject expertise would not have been as insightful.

Stage 2: Use the features and researchers' experience of qualifications to write items about features. Add instructions to the items to form the instrument. Reduce the length of the instrument in preparation for piloting.

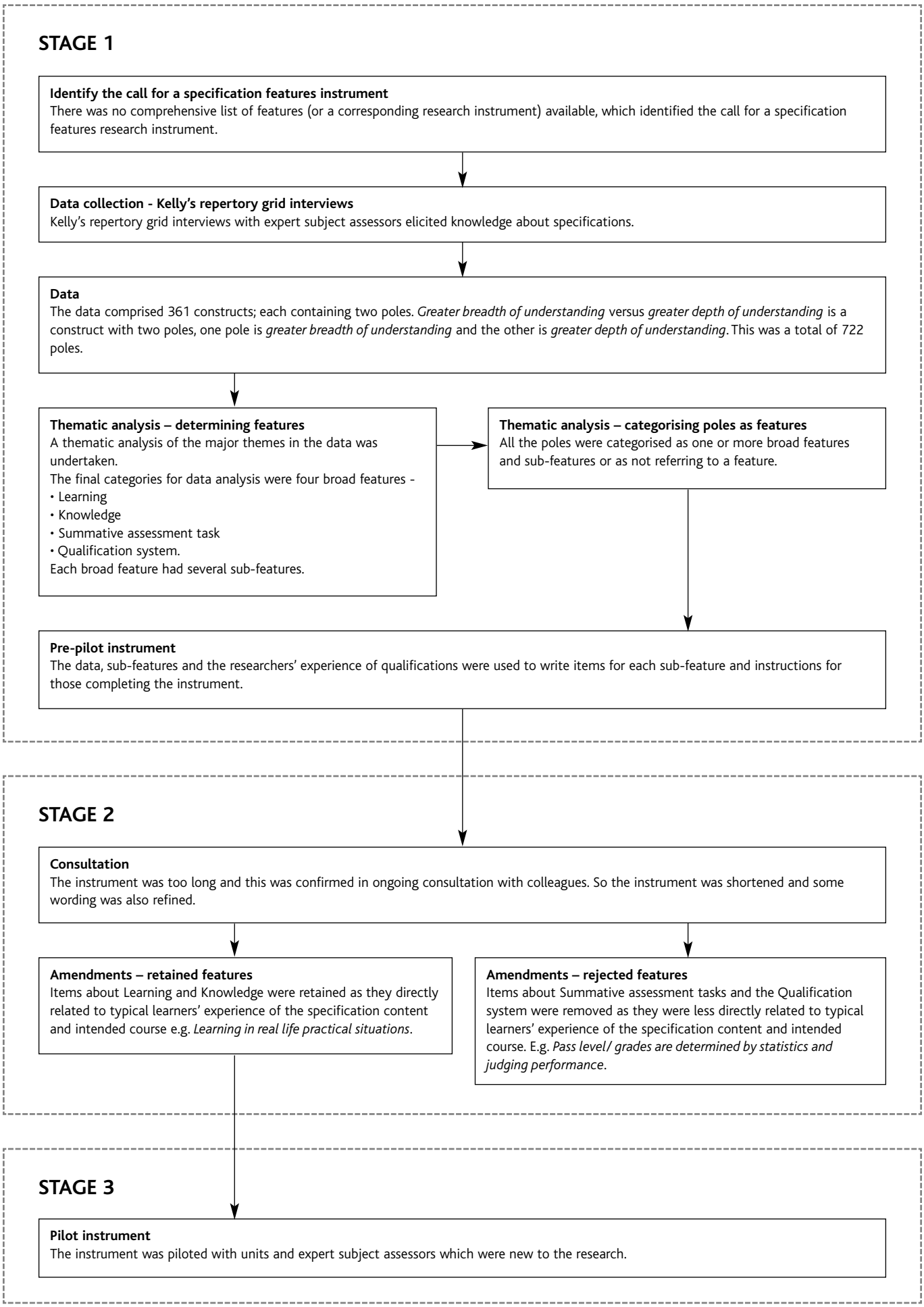
Stage 3: Pilot the instrument.

Figure 1 also summarises the process.

This strategy of combining Kelly's repertory grid interviewing and qualitative analysis is an established strategy for instrument

1. Definitions are also provided by www.examofficers.org.uk/jargon-buster, https://examiners.aqa.org.uk/eap/eap-login/Glossary.action#def27 and http://www.ofqual.gov.uk/help-and-support/94-articles/34-161-glossary#_S; all accessed on 8 December 2010
2. Diplomas are composite qualifications, made up of several free standing qualifications. Some compulsory parts are PL units, Functional Skills in English, Mathematics and ICT. In other areas of the Diploma learners have more choice about which units to study. The Diplomas were first awarded in 2009. Ertl and Stasz (2010) explain that Diplomas are sometimes incorrectly mistaken for VQs.
3. The information is this paragraph is sourced from Ofqual (2010).

Figure 1: Instrument development process



development, see for example Lambert *et al*. (1997), and Edwards and Adams (2002, 2003).

Stage 1: Research to identify features

Method

Interview technique

Kelly's repertory grid (KRG) technique is a well-established research technique for gaining insights into how people view their world. There are several texts about KRG, such as Fransella *et al*. (2004), Beail (1985) and Easterby–Smith (1980).

Easterby–Smith (1980, 4) writes that:

> *A full repertory grid contains three components: "elements", which define the material upon which the grid will be based; "constructs", which are the ways that the subject is grouping and differentiating between the elements; and a "linking mechanism" which can show how each element is being assessed on each construct*.

The KRG interviewing in the present paper is concerned only with the elements (units from a variety of qualifications) and constructs (expert subject assessors' views of how units are similar and different). Repertory grid interview questions generally ask participants how two elements are similar to and different from a third element. This method was applied in the development of this instrument.

Elements

Three subjects were included in the research: Creative and media; Engineering; and Society, health & development. For each subject cognate units were selected as follows:

- Creative and media: one GQ unit, one PL unit and one VRQ unit.
- Engineering: two GQ units, one PL unit and one VRQ unit.
- Society, health and development: one GQ unit, one PL unit, two VQ units and two VRQ units.

There were four GCSE units, three OCR National units, one unit from an OCR Certificate, and two NVQ units. Thus this choice of units included a variety of types of qualification.

Extracts from the specifications were used rather than the whole specification. The extracts contained the following information:

- Aims and objectives
- Unit content
- Grade or performance descriptors (if applicable)
- Assessment and qualification structure
- Information about guided learning hours and length of assessments
- Teaching arrangements.

The removal of any additional information was intended to facilitate and focus the process of eliciting views about the interview topic rather than observations of descriptive differences, such as, variations in specification document layout.

Expert subject assessors

Four expert subject assessors from each subject and with a senior level of responsibility for at least one of the units/qualifications participated. Due to this broad experience the expert subject assessors were well placed to discuss the specifications.

Table 1: Examples of constructs

Number of construct	*Subject*	*Pole 1* *Similarity between units*	*Pole 2* *How a unit(s) was different*
Construct 1	Creative and media	Candidates have ongoing assessment opportunities, so are under less pressure.	Candidates have a single assessment opportunity, so must perform under high pressure.
Construct 2	Engineering	The learner is on their own in the exam.	The learner can ask the presenter for prompts (help) and they can guide the learner in the assignment but not give an answer to the task.
Construct 3	Society, health and development	Some learning and assessment is carried out in unfamiliar situations.	Learning and assessment is mostly carried out in familiar situations.

Interviews

Four interviewers underwent an interview practice and standardisation process prior to interviewing expert subject assessors. These practice and standardisation interviews were undertaken face to face as well as over the telephone with colleagues from Cambridge Assessment.

Two interviewers conducted each interview with each of the expert subject assessors. Prior to the interviews the expert subject assessors were briefed on the task.

To ensure a full complement of expert subject assessors three were interviewed by telephone. The other nine were interviewed face to face.

Secondary analysis – data management and categorising

Strategy and summary

A total of 361 constructs were elicited in the KRG interviews (see Table 1 for examples). All 361 elicited constructs (722 poles) were analysed. Each pole was categorised by a joint panel decision. The panel consisted of three of the four interviewers.

Features

Generally, in thematic analysis researchers read and re-read the data carefully and identify the main themes. The themes are used as categories for the data. The data can be categorised and re-categorised until the researchers arrive at the best categories. Finally, the data in each category are summarised and the relationships between categories are discussed. For further details about thematic analysis see Fereday and Muir-Cochrane (2006), Yawn (2003) and Warner and Griffiths (2006).

Thematic analysis was applied to the KRG data. Reading the poles indicated that some content referred to the teaching and learning situation. For instance, some of the poles contained information about the amount of guidance the learner received from the teacher, or contained information about the breadth of knowledge covered by the specification. Several categories were devised and poles were assigned to categories. However, continual reading and categorising identified further content in the poles. Therefore the categories were revised and the poles re-categorised. Each category was a feature of specifications.

The final categories comprised four broad features each with sub-features (Table 2).

Table 2: Features and sub-features with examples of associated poles

Feature	*Sub-feature*	*Example of pole*	*Subject*
Learning	Level/type of support	Giving someone an essay to write without help or support	Society, health and development
	Familiar and unfamiliar situations	Candidates are taken out of their comfort zone, this develops their personality	Engineering
	Level of interaction/who the learner interacts with	Candidates are required to give, receive and act on peer feedback as part of the group process	Creative and media
	Context of learning, i.e. classroom/practical/ vocational /real life	Closer relationship to business and commercial sector	Creative and media
	Predictability of the situation, how much control the learner has	Candidates have the time and flexibility to experiment, remedy or change direction	Creative and media
	Procedural / declarative knowledge	Technical aspects of setting up for an event (stage management)	Creative and media
	Self-organisation	Organising from own perspective and perspective of others involved	Society, health and development
Knowledge	Breadth/depth of knowledge	Broad knowledge required	Creative and media
	Prior knowledge required for the learning programme	Knowledge base required, needs KS3 as preparation for course	Engineering
	Concrete knowledge	Candidates are required to demonstrate spatial ability	Engineering
	Abstract knowledge	Candidates are required to have knowledge of values	Society, health and development
Summative assessment task	Level/type of support	The learner can ask the presenter for prompts (help) and they can guide the learner in the assignment but not give an answer to the task	Engineering
	Number of summative assessment opportunities	No ability to upgrade evidence. Only change through retakes	Society, health and development
	Familiar and unfamiliar situations	Learning and assessment is mostly carried out in familiar situations	Society, health and development
	Level of interaction/who the learner interacts with	Group works together for whole of examined time	Creative and media
	Context of assessment, i.e. classroom/practical/ vocational/real life	Controlled assessment – all done in classroom (except preparation)	Creative and media
	Predictability of the situation, how much control the learner has	Candidates have to effectively manage and organise their time in order to complete the assessed tasks in a short time period	Creative and media
	Procedural/declarative knowledge	Requires learners to assimilate knowledge in order to produce portfolio evidence	Society, health and development
Qualification system	Available certification outcomes	Range of grades between A–C (Dip), National (A–C), GCSE (A–G)	Society, health and development
	Referencing style	Learning outcomes, assessment criteria and exemplifications, and grade descriptors are provided	Engineering
	Mode of evidence	Blend of written evidence and portfolio evidence (could be presentation etc.)	Society, health and development
	Mode of assessment	Model assignment produced by board or tutor written assignment	Society, health and development
	Who makes assessment judgements	Examiner assessed	Creative and media

Notes:

Level/type of support refers to "Level/type of support (e.g. independent performance/unstructured task versus help provided/structured task)".

Predictability of the situation, how much control the learner has refers to:

- Predictability of the situation
- How much control the learner has/time/time pressure/ deadlines and the flexibility of time and deadlines
- Pressured decision making versus on going decision making

Dealing with uncertainty versus responding to routine situations.

Knowledge refers to "Characteristics of the knowledge learners are exposed to".

Summative assessment task refers to "Summative assessment task and gathering evidence of achievements for a portfolio/equivalent".

Available certification outcomes is short for "Available certification outcomes-pass or fail/range of grades (or equivalent) available/range of levels available"

Referencing style is short for "Referencing style – Criterion referenced/cohort referenced/compensation/norm referenced/ descriptor referenced (judgement of best fit)/hurdles"

Mode of evidence is short for "Mode of evidence – response to a standardised test or task (such as a script)/portfolio/verbal evidence/written evidence/another form of performance evidence"

Mode of assessment is short for "Mode of assessing – standardised test or task (such as an examination)/verbal questioning/task determined by the candidate/task determined by the assessor/teacher (but not a standardised task)"

Who makes summative assessment judgements is short for "Who makes assessment judgements – external examiner/ internal assessor"

Judgement process

First, each pole was categorised as belonging to none, one, or more of the following features:

- Learning
- Knowledge
- Summative assessment task
- Qualification system.

Secondly, each pole was categorised with one or more of the sub-features in.

Consensus was reached through panel discussion. Once all the data were categorised into features and sub-features the panel revisited the data to confirm the decisions.

Table 2 contains examples of poles, the features and sub-features they were assigned to and the subject in which the pole was situated.

Findings

This section considers the results of the analysis.

Table 3 presents the frequency of expert subject assessors whose KRG data included one or more poles assigned to each sub-feature. The data are also organised by subject and type of qualification. It can be seen that most features related to all three subjects and all four qualification types.

Some poles did not refer to features but referred to topics such as the stakeholders involved in writing the specification. Therefore they were excluded from the instrument development process.

Stage 2: Using research evidence to write a features instrument

The next stage in development was to write the features instrument from the research results.

The panel wrote items for each sub-feature and instructions for those completing the features instrument. The data, the sub-features and the panel's experience of qualifications were used in this process. The features and items are provided in Table 4. The items were about features.

Throughout the process of instrument development colleagues were consulted. The ongoing consultation suggested that the instrument was too long and that some wording needed refining. To shorten the instrument the items about the features 'Summative assessment task' and 'Qualification system' were removed as they were less directly related to typical learners' experience of the specification content and intended course, for example, typical learners might not know whether *'Pass level/ grades are determined by statistics and judging performance'*. The features 'Learning' and 'Knowledge' were retained as they were directly related to typical learners' experience of the specification content and intended course, for example, *Learning in real life practical situations*. It was not possible to reduce the length of the instrument by integrating similar items into one item as each item was about a different topic.

The next stage was piloting the instrument.

Table 3: Frequency of expert subject assessors whose KRG data included the presence of poles assigned to each sub-feature

Features		*Creative and media*			*Engineering*			*Society, health & development*			
		GQ	*PL*	*VRQ*	*GQ*	*PL*	*VRQ*	*GQ*	*PL*	*VQ*	*VRQ*
Learning	Level/type of support	3	4	3	3	3	3	3	3	4	3
	Familiar and unfamiliar situations	1	1	1	2	1	2	1	1	1	1
	Level of interaction/who the learner interacts with	2	4	1	2	1	3	3	3	3	1
	Context of the assessment i.e. classroom/practical/vocational/real life	4	4	4	3	3	4	3	4	3	2
	Control/time pressure/decision making	3	4	4	2	2	1	2	2	3	1
	Procedural/declarative knowledge	4	4	4	4	2	3	2	1	3	2
	Self organising versus set structure	1	1	1	1	1	1	2	2	1	1
Knowledge	Breadth and depth	4	4	3	4	3	4	4	4	4	3
	Prior knowledge required for the learning programme	1	0	0	2	1	0	0	1	0	0
	Concrete	4	4	4	4	4	4	3	4	4	3
	Abstract	0	0	1	0	0	0	2	2	2	3
Qualification system	Available certification outcomes	0	0	1	0	0	0	2	2	2	2
	Referencing style	0	1	1	0	1	0	0	0	0	0
	Mode of evidence	2	2	1	0	0	0	0	0	0	1
	Mode of assessment	0	0	0	0	0	0	1	0	0	1
	Who makes summative assessment judgements	1	1	0	0	0	0	0	0	0	0
Summative assessment task	Level/type of support	0	0	1	1	1	0	2	2	1	2
	Number of summative assessment opportunities	1	1	1	0	0	0	2	1	0	1
	Familiar and unfamiliar situations	0	0	0	0	0	0	1	1	0	1
	Level of interaction/who the learner interacts with	1	1	0	0	1	0	0	0	0	0
	Context of the assessment i.e. classroom/practical/vocational/real life	2	1	1	2	2	1	0	2	1	0
	Control/time pressure/decision making	2	2	1	4	3	2	2	2	1	2
	Procedural/declarative knowledge	1	0	0	1	1	0	1	1	0	1
	Self organising versus set structure	1	0	0	0	0	0	0	1	0	0

Table 4: Features and resulting items

Feature	*Items*
Learning	
Level/type of support	Learning through independent performance Learning supported through help provided Learning through structured tasks Learning through unstructured tasks
Familiar and unfamiliar situations	Learning in familiar situations Learning in unfamiliar situations
Level of interaction/who the learner interacts with	Learner works individually Learner works in a group Learner interacts with the public Learner interacts with other learners as part of learning
Context of the learning i.e. classroom/practical/ vocational/real life	Learning in the classroom Learning in real life practical situations Learning through situations that simulate real life
Control/time pressure/ decision making	Learning is time pressured Learning is not time pressured Learning has deadline Learner has control over the learning situation Learner has limited or no control over the learning situation
Procedural/declarative knowledge	Learner develops procedural knowledge Learner develops factual knowledge
Self organising versus set structure	Learner organises their own time to complete task Learner works to an imposed timetable
Knowledge	
Breadth and depth	Learner develops broad knowledge Learner develops narrow range of knowledge Learner develops in-depth knowledge Learner develops basic knowledge Learner assessed on broad knowledge Learner assessed on narrow range of knowledge Learner assessed on in-depth knowledge Learner assessed on basic knowledge
Prior knowledge required for the learning programme	Prior knowledge required for learning No prior knowledge required for learning
Concrete	Learner develops concrete knowledge Learner assessed on concrete knowledge
Abstract	Learner develops general understanding and awareness Learner assessed on general understanding and awareness Learner develops abstract knowledge Learner assessed on abstract knowledge
Qualification system	
Available certification outcomes	Certification outcomes are pass and no pass (or equivalents) Certification outcomes are a series of grades (or equivalents)
Referencing style	Pass level/grades are determined by criteria which learners must meet Pass level/grades are determined by statistics and judging performance Pass level/grades are determined by statistics only Pass level/grades work on a principle of compensation (strengths are rewarded and no credit is lost for weaknesses) Pass level/grades include hurdles (one aspect of learners' performance must meet a particular criterion but the rest of the performance is judged differently) Applying a judgement of best fit
Mode of evidence	Learners can be assessed on their written evidence Assessment includes another form of evidence
Mode of assessment	All learners are assessed using the same task/exam Assessment tasks vary with centres/learners All learners are assessed on their portfolios Assessment includes verbal questioning and responses The assessment task is determined by the learner The assessment task is determined by an assessor
Who makes summative assessment judgements	An external assessor makes assessment judgements An internal assessor makes assessment judgements
Summative assessment task	
Level/type of support	Assessed on independent performance Assessment is supported through help provided Assessed on structured tasks Assessed on unstructured tasks
Number of summative assessment opportunities	Unlimited assessment opportunities Limited assessment opportunities
Familiar and unfamiliar situations	Assessment in familiar situations Assessment in unfamiliar situations
Level of interaction/who the learner interacts with	Learner produces individual work for assessment Learner works in a group for assessment Learner interacts with the public as part of assessment Learner interacts with other learners as part of assessment
Context of the assessment i.e. classroom/practical/ vocational/real life	Assessment in the classroom Assessment in real life practical situations Assessment in situations that simulate real life
Control/time pressure/ decision making	Assessment is time pressured Assessment is not time pressured Assessment has deadlines Assessment has no deadlines Learner has control over the assessment situation Learner has no control over the assessment situation
Procedural/declarative knowledge	Learner assessed on procedural knowledge Learner assessed on factual knowledge
Self organising versus set structure	Learner organises their own time for assessment Learner works to an imposed timetable

Stage 3: Pilot of the features instrument

The purpose of the features research instrument is to compare the characteristics of knowledge and learning associated with cognate units from different types of qualifications, such as vocational and general qualifications. Therefore, the following research question is posed:

> *Is the research instrument appropriate for use in research studies? (i.e. do research results from the research instrument compare between the different types of units?)*

It was considered useful to also investigate whether the results from the instrument compare between units of the same type, and this became a subsidiary research question.

Method

Units

Four cognate level two units in Health were selected, two from an NVQ, one from a current GCSE and one from a legacy GCSE. For the purposes of this article the units were called NVQ1, NVQ2, GCSE1 and GCSE2. None of the units had been used in earlier parts of the research.

Expert subject assessors

Four expert subject assessors were recruited. The criteria for selection were that they:

- were a Team Leader, Assistant External Verifier or above for one of the qualifications
- were recommended by OCR
- did not participate in earlier parts of the research.

The first two criteria are used in some other comparability studies. The expert subject assessors were paid volunteers.

Materials

The expert subject assessors were provided with the instrument (see Appendix 1) and specification extracts.

Procedure

The expert subject assessors completed the instrument remotely and individually, then returned it to the Research Division. The data collection took place in December 2010.

Table 5: Frequency of responses

		NVQ1	*GCSE1*	*NVQ2*	*GCSE2*
1	Learning through independent performance	4	3	4	4
2	Learning supported through help provided	4	4	4	4
3	Learning through structured tasks	2	4	2	4
4	Learning through unstructured tasks	4	1	4	1
5	Learning in familiar situations	2	2	3	2
6	Learning in unfamiliar situations	4	2	3	3
7	Learner works individually	3	3	4	3
8	Learner works in a group	3	2	3	2
9	Learner interacts with the public	3	1	3	2
10	Learner interacts with other learners as part of learning	3	3	3	4
11	Learning in the classroom	2	4	2	4
12	Learning in real life practical situations	4	2	3	3
13	Learning through situations that simulate real life	4	4	2	4
14	Learning is time-pressured	1	3	2	2
15	Learning is not time-pressured	4	2	4	3
16	Learning has deadlines	1	4	3	2
17	Learning has no deadlines	3	1	1	2
18	Learner has control over the learning situation	3	1	3	2
19	Learner has limited or no control over the learning situation	1	3	1	3
20	Learner develops procedural knowledge	3	4	3	2
21	Learner develops factual knowledge	4	4	4	4
22	Learner organises their own time to complete task	3	1	3	1
23	Learner works to an imposed timetable	1	3	1	3
24	Learner develops broad knowledge	3	4	3	2
25	Learner develops narrow range of knowledge	1	0	1	2
26	Learner develops in-depth knowledge	2	3	3	2
27	Learner develops basic knowledge	3	2	2	3
28	Prior knowledge required for learning	2	1	2	2
29	No prior knowledge required for learning	4	3	3	3
30	Learner develops concrete knowledge	4	4	4	4
31	Learner develops general understanding and awareness	2	4	2	4
32	Learner develops abstract knowledge	3	2	4	3

Findings

Do research results from the research instrument compare between the different types of units?

The features are relevant to some units beyond those used in Stage 1 of the development. As Table 5 shows at least one expert subject assessor thought each feature was relevant to each unit. The exception was one feature (25) and one unit (GCSE1).

The results can be used to identify similarities between units. For instance, Table 5 shows expert subject assessors agreed the following items were common to all units:

- (2) Learning supported through help provided
- (21) Learner develops factual knowledge
- (30) Learner develops concrete knowledge

The results for all three items above show comparisons can be made between the features of cognate units of the same type (i.e. NVQ1 and NVQ2; GCSE1 and GCSE2) or different types (e.g. NVQ1/NVQ2 and GCSE1/GCSE2).

The results can be used to identify differences between units. An example is that all four expert subject assessors agreed (12) *Learning in real life practical situations* was relevant to NVQ1 but there was less agreement on whether this feature was relevant to each of the other units (Table 5). This example illustrates that comparisons can be made between the features of cognate units of the same type (i.e. NVQ1 and NVQ2) or different types (e.g. NVQ1 and GCSE1/GCSE2).

Therefore, as hoped, the research instrument highlighted the similarities and differences between units. This was the case for units of the same type and different types.

Conclusion

This article describes the development of a features instrument. The instrument was intended to:

- Compare features of cognate units from different types of qualifications
- Be suitable for re-use in various subjects.

The instrument (Appendix 1) is considered appropriate because it is based on expert subject assessors' views. The instrument presents a list of specification features derived from the perspective of expert subject assessors. The list of specification features is given in the form of items. Their expert views contextualised the specifications in the appropriate subject assessment community to formulate constructs. A document analysis of the specifications by researchers who do not have the subject expertise would not have been as insightful. That the expert subject assessors represented three subjects and different types of qualifications adds credibility to the resulting instrument. Additionally, the features and instrument read as if they apply to all types of qualifications in the research and various subjects beyond the three studied here. The pilot study indicated that salient features vary somewhat between units. Therefore, as hoped, the research instrument highlights similarities and differences between units, and this is the case for units of the same type and different types.

References

Beail, N. (1985). *Repertory Grid Technique and Personal Constructs Applications in Clinical and Educational settings*. Kent: Croom Helm Ltd.

Easterby-Smith, M. (1980). The design, analysis and interpretation of repertory grids. *International Journal of Man-Machine Studies*, **13**, 1, 3–24.

Edwards, E. & Adams, R. (2003). *A comparability study in GCE Advanced level geography including the Scottish Advanced Higher grade examination*. A study based on the summer 2002 examination. Organised by the Welsh Joint Education Committee on behalf of the Joint Council for General Qualifications.

Edwards, E. & Adams, R. (2002). *A comparability study in GCE AS geography including parts of the Scottish Higher grade examination*. A study based on the summer 2001 examination. Organised by the Welsh Joint Education Committee on behalf of the Joint Council for General Qualifications.

Ertl, H, & Stasz, C. (2010). Employing an 'employer-led' design? An evaluation of the development of Diplomas. *Journal of Education and Work*, **23**, 4, 301–317.

Fereday, J. & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of qualitative methods*, **5**, 1, 80–92. http://ejournals.library.ualberta.ca/index.php/IJQM/article/viewArticle/4411 [Accessed 4th October 2010].

Fransella, F. Bell, R. & Bannister, D. (2004). *A Manual for Repertory Grid Technique*. Second Edition. Chichester: John Wiley and Sons Ltd.

Lambert, R., Kirksey, M. & McCarthy, C. (1997). *The repertory grid as a qualitative interviewing technique for use in survey development*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, March 24–28, ED409367.

Newton, P. (2007). Contextualising the comparability of examination standards, In: P. Newton, J. Baird, H. Goldstein, H. Patrick, P. Tymms (Eds.). *Techniques for monitoring the comparability of examination standards*. 9–42. London: Qualifications and Curriculum Authority.

Novaković, N. & Greatorex, J. (2011). Comparing the demand of syllabus content in the context of vocational qualifications: literature, theory and method. *Research Matters: A Cambridge Assessment Publication*, **11**, 25–32.

Ofqual (2010). http://www.ofqual.gov.uk/qualification-and-assessment-framework/89-articles/145-explaining-the-qualifications-and-credit-framework. The website was last updated in 16th June 2010 [Accessed 2nd December 2010]

QCA (2007a). GCSE (short course) citizenship studies comparability study QCA/07/3089 http://ofqual.gov.uk/files/QCA-07-3089_GCSE_citizenship_mar07.pdf [Accessed 16th September 2010].
QCA (2007b). GCSE French comparability study QCA/07/3098 http://www.all-london.org.uk/Resources/severe_grading/QCA-07-3098_GCSE_French_mar07.pdf [Accessed 16th September 2010].

Warner, R. & Griffiths, M. D. (2006). A qualitative thematic analysis of exercise addiction: An exploratory study. *International Journal of Mental Health Addiction*, **4**, 13–26.

Yawn, B. P. (2003). The impact of childhood asthma on daily life of the family – a qualitative study using recurrent thematic analysis. *Primary Care Respiratory Journal*, **12**, 3, 82–85.

APPENDIX 1 — Features research instrument

Instructions

This research instrument was developed to systematically list features of different level 2 specifications and identify which features are present in different specifications.

Listed in the instrument are features of learning and knowledge which some specifications intend typical level 2 learners to experience.

Please read the list carefully and tick the boxes to indicate the relevant features. These features may be explicit in the specifications or implicit and part of an underpinning ethos.

If you find there are additional features intended by the specification which are not in the list, please add them in under 'other' at the end of the instrument.

Please ensure you have familiarised yourself with the specifications before starting this task.

Feature	Indicate if feature is present in			
Questions 1 to 19 are about **Learning**	NVQ1	GCSE1	NVQ2	GCSE2
1 Learning through independent performance				
2 Learning supported through help provided				
3 Learning through structured tasks				
4 Learning through unstructured tasks				
5 Learning in familiar situations				
6 Learning in unfamiliar situations				
7 Learner works individually				
8 Learner works in a group				
9 Learner interacts with the public				
10 Learner interacts with other learners as part of learning				
11 Learning in the classroom				
12 Learning in real life practical situations				
13 Learning through situations that simulate real life				
14 Learning is time-pressured				
15 Learning is not time-pressured				
16 Learning has deadlines				
17 Learning has no deadlines				
18 Learner has control over the learning situation				
19 Learner has limited or no control over the learning situation				

Feature Questions 20 to 32 are about **Knowledge**	Indicate if feature is present in			
	NVQ1	GCSE1	NVQ2	GCSE2
20 Learner develops procedural knowledge				
21 Learner develops factual knowledge				
22 Learner organises their own time to complete task				
23 Learner works to an imposed timetable				
24 Learner develops broad knowledge				
25 Learner develops narrow range of knowledge				
26 Learner develops in-depth knowledge				
27 Learner develops basic knowledge				
28 Prior knowledge required for learning				
29 No prior knowledge required for learning				
30 Learner develops concrete knowledge				
31 Learner develops general understanding and awareness				
32 Learner develops abstract knowledge				

Other features Use this space to add any features intended by the specification which you feel have not been covered.	Indicate if feature is present in			
	NVQ1	GCSE1	NVQ2	GCSE2

To request permission to use or adapt the features research instrument write to Jackie Greatorex, Research Division, Cambridge Assessment, 1 Regent Street, Cambridge CB1 2EU.

APPENDIX 2 — Searches of the national database of accredited qualifications (NDAQ)

Search options			*Results*		
Subject	*Type*	*Level*	*Matches*	*NDAQ Types*	*Awarding bodies*
All	All	All	11258	EL (Entry Level) ESOL (English for Speakers of Other Languages), FS NQF (Functional Skills National Qualifications Framework) GCE (General Certificate of Education) GCE AS (General Certificate of Education Advanced Subsidiary) GCSE (General Certificate of Education) HL (Higher Level Qualifications) NVQ (National Vocational Qualification) OG (Other General Qualification) OQ (Occupational Qualification) PL (Principal Learning) PROJ (Project) QCF (Qualification and Credit Framework) VRQ (Vocationally-Related Qualification)	139
Art	-	2	230	GCSE, NVQ, OG, OQ, QCF, VRQ	21
Business	-	2	162	ESOL , GCSE, NVQ, OG, OQ, PL, VRQ	25

Notes:
1. The National Database of Accredited Qualifications (NDAQ) held details of qualifications that are accredited by the qualification regulators in England (Ofqual), Wales (DCELLS) and Northern Ireland (CCEA) http://www.accreditedqualifications.org.uk /index.aspx:
2. All the searches were restricted to current qualifications and qualifications offered in English language only.
3. Awarding bodies is used here to refer to awarding bodes and collaborations between awarding bodies.

Statistical Reports

The Statistics Team Research Division

Four new reports have been added to the 'Statistics Report Series' on the Cambridge Assessment website since the publication of Issue 11 of *Research Matters*. The reports in this series provide statistical summaries of various aspects of the English examination system such as trends in pupil attainment, qualifications choice and subject uptake and provision at school.

The following reports, produced using national-level examination data, are available at http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Statistical_Reports:

- Statistics Report Series No. 25: Uptake of ICT and computing qualifications in schools in England 2007–2009
- Statistics Report Series No. 26: The accuracy of forecast grades for OCR A levels
- Statistics Report Series No. 27: Provision of GCE A level subjects 2010
- Statistics Report Series No. 28: Uptake of GCE A level subjects 2010

Research News

Conferences and seminars

Parliamentary Research Enquiry

Cambridge Assessment's seventh Parliamentary Research Enquiry seminar took place in the House of Commons in January. The event, chaired by Graham Stuart MP, Chair of the Education Select Committee, was jointly organised by Cambridge Assessment, the University of Cambridge's international exams group, and the University's Centre for Science and Policy.

Experts in neuroscience, psychology and education emphasised the importance of motivational and contextual influences, and the importance of active, directed learning in ensuring that a child's potential is realised. They agreed that neuroscience needs to have a bigger impact on policymakers than at present, and that the discourse needs to shift to a focus on children and learning, both in terms of cognitive and emotional development.

Speakers included Usha Goswami, Professor of Cognitive Developmental Neuroscience at the University of Cambridge, Robert Burden, Emeritus Professor of Applied Educational Psychology at the University of Exeter, and Trevor Robbins, Professor of Cognitive Neuroscience at the University of Cambridge.

Discussion spanned a number of topics including: the role that language plays in the early years; the importance of structures and support systems; whether the cognitive learning processes had determined the break points in education; and the age at which children should start school.

Cambridge Assessment's Parliamentary Research Enquiry series is designed to bring together a wide range of professionals in education to look at 'big picture' topics and enable policy makers to access the knowledge of leading experts.

Podcasts of the event are available at www.cambridgeassessment.org.uk

5th UK Rasch User Group Meeting

In January Tom Bramley and Beth Black attended the 5th UK Rasch User Group meeting at the CEM Centre in Durham. The UK Rasch Users Group Meeting provides a forum for Rasch enthusiasts working in different fields to get together to share ideas and present research. The purpose of the group is to offer advice, support and encouragement to anyone interested in the Rasch model.

The Rasch Day itself was followed by a one day workshop on 'The R environment and estimation of the Rasch Model', tutored by Tima Croudace of the University of Cambridge and Jan Boehnke of the Universituy of Trier.

American Educational Research Association (AERA)

The AERA annual conference took place in New Orleans in April with the theme of 'Inciting the social imagination: Education research for the public good'.

Irenka Suto and Victoria Crisp were invited to present two papers as part of a collaborative symposium with American colleagues on 'Rater cognition and its importance for score validity: Global perspectives and findings'.

Victoria gave a paper on 'An investigation of rater cognition in the assessment of projects', while Irenka presented on 'A critical review of research methods used to explore rater cognition'.

Publications

The following articles have been published since Issue 11 of *Research Matters*:

Bramley, T. (2010). What can be inferred about classification accuracy from classification consistency? *Educational Research*, **52**, 3, 325–330.

Emery, J.L., Bell, J.F. & Vidal Rodeiro, C.L (2011). The BioMedical Admissions Test for medical student selection: Issues of Fairness and Bias. *Medical Teacher*, **33**, 1, 62–71.

Johnson, M., Nádas, R., & Bell, J.F.(2010). Marking essays on screen: an investigation into the reliability of marking extended subjective texts. *British Journal of Educational Technology,* **41**, 5, 814–826.

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

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

ISSN: 1755–6031