Comparability of national tests over time: a project and its impact

Alf Massey Head of Evaluation and Validation Unit

The Comparability Over Time (CoT) Project was commissioned by the Qualifications and Curriculum Authority (QCA) and the research team, based at UCLES, began work in April 1999. The project investigated the stability of national test standards at all key stages and in all subjects and a final report (Massey, Green, Dexter & Hamnett, 2003) was made public by the QCA in December 2003. Since then the project's findings have attracted (and continue to attract) much comment. National test standards are of considerable public interest, not least because of the political prominence these tests have been accorded, including government claims that the huge improvements in results since tests were introduced in the mid-1990s stem from the plethora of recent educational policy initiatives.

Why comparability over time has proved elusive

The search for comparability over time in large scale assessments in some ways resembles that for the Holy Grail: it has a somewhat mythological quality; it is difficult to find and even many of those who should know better are uncertain what it might look like. Comparability over time is thus famously difficult to investigate (Goldstein, 1983; Newton, 1997) and even the language we use to talk about educational standards is often ambiguous (Massey, 1994).

As is the case with UK public examinations, new versions of England's national tests are set annually. Inevitably difficulty will vary somewhat from year to year; posing an annual standard setting conundrum. Each year we must decide which mark merits the award of each level (i.e. set level thresholds) so as to 'maintain standards'. Yet even in this context, where whole national cohorts take successive versions of a test, it is difficult – probably impossible – to provide an absolutely sound methodology to guarantee equivalence over time. Decisions must disentangle the effects of changes in the quality of teaching and learning (themselves affected by national and local policy initiatives and resource issues etc.) from the effects of variations in things like:

- the curriculum itself, both formal and informal (i.e. how teachers choose to implement it);
- test content and questions and, hence, 'absolute' difficulty;
- the calibre of samples taking given test forms for trial or equating purposes;
- the format of assessment used – including deliberate enhancements and changes in style;
- wider aspects of culture and social expectations that influence children's responses.

Different methodologies

Judgemental comparison of the equivalence or level of demand of the question papers themselves is endlessly fascinating, but research teaches us very clearly that teachers or other experts cannot predict the likelihood of students answering correctly accurately enough to allow test thresholds to be set, safely, on this basis. Judgemental comparisons involving pupils' work are more feasible and have the advantage of enabling comparisons across diverse assessments and contexts, but they are rather blunt instruments, probably only able to detect gross changes. This limits reliance on them in standard setting (Cresswell, 2000). But such judgements remain of interest. They are often valued by decision takers and other interested parties and there are systematic approaches in development which show some promise (Bramley, 2005a). Christie and Forrest (1981) used expert judgements to compare GCE AL standards over a ten year period and their work showed how much the criteria implicit in examinations change and hence make comparisons more difficult. Analytic comparisons (e.g. Massey & Elliott, 1996) are rare, partly for lack of substantial archives of scripts from the past, but have something to offer in investigating the variations in the ways in which achievement is demonstrated over longer periods of time.

Indirect comparisons can provide interesting data, using reference measures (which remain the same from year to year and may take the form of prior achievement or other 'existing' measures, such as scores on standardised tests) taken by children in successive cohorts to estimate their relative ability/achievement. Do children with similar reference test scores fare equally well on each year's test? The CoT Project made considerable use of this approach, gathering extensive data thanks to the co-operation of Local Education Authorities (LEAs). But the effects of curricular change on what is learned and assessed are considerable, especially over intervals of more than a few years, and those preferring different outcomes can disagree about how such factors might affect test performance. Argument about the varying relevance of the common measure to assessments in different years often makes conclusions based on this approach indicative rather than conclusive (Newbould and Massey, 1979).

Direct experimental comparisons seem an obvious method, but have not previously been used on a large scale in the context of Britain's curriculum driven performance assessment systems because of both high costs and the rate of curricular change, which quickly makes 'past' tests out of date for today's children and hence invalidates comparisons. It is worth noting that it was not until 1996, the base year for our study, that the national test system approached stability. Comparisons like ours could not have been commissioned earlier and QCA should take credit for their promptness in initiating this work when they did. Quasi-experiments, equating next year's test forms to those taken operationally by today's children, have been part of national tests' standard setting

process since their inception. But these are bedevilled by pupils' varying motivation levels when taking real and trial tests. Performance appears to fluctuate unpredictably – perhaps according to variations in school culture which help determine how seriously pupils take non-operational tests. The CoT Project was fortunate to enjoy the resources and opportunity to use medium-term direct experimental comparisons as the major basis for comparisons, thus avoiding such problems.

The Comparability Over Time Project

The project's work involved two main strands of quantitative research, supported by two qualitative approaches:

Medium-term direct experimental comparisons

Randomly assigned groups of children (in Northern Ireland – NI) took either 1996 or later versions of national tests in all subjects at all key stages, in the course of three (annual) phases of experimental testing between 1999 and 2001, as shown below.

- KS1 Reading Comprehension 1996 v 1999
- KS1 Mathematics 1996 v 2000
- KS2 English 1996 v 1999 and 1996 v 2000
- KS2 Mathematics 1996 v 1999
- KS2 Science 1996 v 2001
- KS3 English 1996 v 2001
- KS3 Mathematics 1996 v 2000
- KS3 Science 1996 v 2001

These intervals were short enough that the tests being compared were not so dissimilar that comparisons were invalidated and long enough for detectable effects to have appeared.

Children in NI were used so that they had not seen either the current or previous versions of these tests beforehand. In all 11,762 children from 184 schools were involved – providing large enough groups (circa 1,000 for each experimental comparison – well distributed across the range of ability) for sufficiently powerful statistical comparisons. Detailed desk research, fieldwork and/or questionnaires for teachers were used to investigate the validity of our conducting research with children following the slightly different curriculum in NI; enabling some (largely minor) issues to be identified, ready to be considered alongside the empirical data when reaching conclusions. Teachers from the schools involved considered these tests to be appropriate for their pupils and whilst they had not been prepared for them as directly as their English counterparts, any consequent reduction in overall performance was irrelevant to our purpose – that of comparing standards across versions rather than estimating how well the children themselves could perform.

For each subject at each key stage, two experimental groups were formed, to take the 1996 or the later version of the test respectively, using spiral quasi-random assignment. Alternate boys and girls on the school or class register were assigned to each form of the test, to minimise gender, school/teaching group and neighbourhood effects by distributing them evenly. However supplementary data concerning performance on relevant NI national assessments (or date of birth in the case of KS1 children for whom NI assessments were unavailable) were also collected, to help compare the equivalence of experimental groups and to use as a control variable if required.

Analyses then used the NI national assessment data (or date of birth) to check the equivalence of the experimental groups, before comparing the distributions of levels each group achieved to establish whether or not any differences were statistically significant. Wherever possible, features of the data available were explored to try to shed light on the origins of disparities observed between test forms.

In subjects/key stages where such variations were observed the data also enabled the project to equate the mark scales of the tests set in different years. If level thresholds set in the later year did not correspond to marks equated to the equivalent thresholds set in the earlier version, it suggests that variations in test results across the years have either under or over estimated the progress made by schools. A varied pattern of results was obtained, across key stages and subjects, as summarised in our conclusions below.

Evidence from LEA's standardised testing in schools

A few of the LEAs in England continue to use standardised tests extensively, often to help allocate resources. In effect these can provide found 'common reference test' data to help compare national test results by looking to see if children with the same standardised test scores from different years obtain equivalent national test results. We canvassed LEAs and gathered data available spanning all or parts of the period 1996–2000 to help cross-validate our experimental evidence.

Different LEAs use different tests and in effect provide a series of case studies. LEA 1 provided data from 29,896 children relating to KS1 English; 29,926 children relating to KS2 English; and 20,788 children relating to KS2 Mathematics – between 1996 and 2000. LEA 2 provided data for 22,985 children regarding KS2 English between 1996 and 1998. LEA 3's data was for 4,772 children concerning KS2 English between 1997 and 1998. LEA 4 provided data for 52,950 children concerning KS1 English between 1996 and 1999. LEA 5's data was for 17,963 children concerning KS2 English and for 17,971 children regarding KS2 Mathematics – between 1998 and 2000. LEA 6 provided data for 13,904 children concerning KS2 English and 15,747 children regarding KS2 Mathematics – between 1996 and 1998.

Evidence across LEAs suggested that (perhaps against the curricular odds, given that schools' attention must have been switching towards national tests as the latter's importance was increasingly recognised by teachers) standardised test scores had risen over the period investigated, in itself suggesting rising standards of teaching and learning.

Despite the very different methodologies and assumptions involved, relating the data on standardised testing to children's national test results provided convincing support for the conclusions reached via the experimental comparisons for the same key stages/subjects, regarding both the size and nature of effects observed. There were also considerable similarities between the data for the various subjects/key stages from different LEAs, so replication (both between methodologies and across these case studies) further bolsters confidence in the project's evidence.

Teachers' judgements about the quality of scripts

A small scale study involved judgemental comparisons (by teachers) of 'representative' 1996 and 1999 KS2 English scripts at key mark points. These teachers' judgements also supported the conclusions reached via experimental comparisons.

Children's perceptions of evolving features in national tests

Children's views about the tests they take are rarely sought. Our project interviewed small samples of children regarding every subject at each key stage (*n* ranging from 12 to 24 in each case), using selected paired

comparisons between materials from 1996 and the more recent test materials. A modified version of Kelly's repertory grid questioning technique was employed to help them verbalise their thoughts and identify salient features and even children as young as seven proved capable of doing so effectively.

Overall the materials sampled from more recent tests tended to be preferred – with some exceptions of course, as earlier versions themselves often contained attractive features. In short, children had appreciated the efforts which had been made to 'improve' the tests over the years by making them more attractive, more user friendly and more accessible.

Such changes will affect both motivation and performance. For instance those which make it easier for children to understand what is required or simplify the ways they respond seem likely to make it easier to demonstrate competence. But such issues raise interesting questions regarding test standards. For instance, should test thresholds be adjusted to compensate for greater user-friendliness? Or should developments which help children show what they can do be seen as a valid means of recognising performance – which should be reflected in improving results? Bramley (2005b) has recently considered the implications of changes in accessibility for measurement models used in the national test context.

Conclusions

- Experimental comparisons suggested that KS1 Reading Comprehension test standards at level 2c were similar in the 1996 and 1999 versions, but those at levels 2a and 2b may have been a mark or two more severe in 1999, so that gains in national test results may under-estimate progress for abler children. The methodologically independent relationships identified between test results and LEA standardised test data were consistent with this.
- Experimental evidence suggested that KS1 Mathematics test standards in 2000 were at least equivalent to those in the 1996 version, and here too levels 2a and 2b were perhaps a mark higher than needed to equate to their equivalents in 1996. Level 3 appeared even more severe in the later version, by around three marks. LEA standardised test data were again consistent with this conclusion and the implication is that improvements posted in national results over this period are in general merited, and may indeed slightly under-estimate progress by abler children.
- Experimental evidence suggested that there were disparities in KS2 English test standards between the 1996 and 2000 versions. It would have been necessary to increase Level 4 and 5 thresholds in the 2000 version by five and seven marks respectively to equate them to 1996 standards. Differences of this order might account for about half the national gains in test results over the period. Given that the writing element in the tests remained almost entirely stable, it can be deduced that the difference was attributable to the Reading element. The project replicated comparisons with the 1996 version in 1999 and 2000 in great detail, which suggested convincingly that the experimental methodology was robust. Relevant data from LEA standardised testing programmes supported the nature and size of the effects observed experimentally here too. In a third independent methodology, albeit in a small-scale qualitative comparison, judgements by teachers asked to compare samples of 1996 and 1999 scripts representing key mark points also concurred with the larger-scale empirical data.
- Experimental data suggested that KS2 Mathematics test standards were similar in the 1996 and 1999 versions, despite the potential for disturbance brought about by the introduction of a mental arithmetic element during this period. Relationships between LEA standardised test data and national test results in successive cohorts supported this too.
- Experimental comparisons indicated that to equate to standards in the 1996 version, thresholds in the 2001 KS2 Science test at levels 2, 3 and 4 would have needed to have been set somewhat higher – by perhaps two marks at levels 2 and 3 and by four marks at level 4. However, 2001's level 5 threshold appeared in line with the standard set in 1996. But changes of this order would account for only a small proportion of the very large gains in KS2 national test results in this subject over the period concerned.
- Because it was impossible to use paper 2 (the Shakespeare element) in experimental comparisons in KS3 English, the methodology had to be adapted here and comparisons were based on predicted test levels generated from experimental comparisons involving only paper 1. Additionally, a small scale re-marking exercise investigated the possibility of 'expectation creep' in contemporary markers' judgements about pupils' writing, compared with those made by markers in 1996. The latter suggested that more demanding marking of writing may have offset slightly lenient thresholds in the reading element detected in the experimental comparisons, leading to the conclusion that no differences could be detected in overall test standards between the 1996 and 2001 versions.
- KS3 Mathematics tests involve a series of attainment-related Tiers, targeted at levels 3–5, 4–6, 5–7 and 6–8 respectively. The experimental comparisons suggested that KS3 Mathematics test standards in the 1996 version appeared more severe than the 2000 version, especially in the lower tiers. Those taking the 2000 version of Tier 3–5 achieved results about half a level better than those taking the 1996 version; in Tiers 4–6 and 5–7 results were about one quarter of a level better on the 2000 version; and in Tier 6–8 achievement was only about one tenth of a level better on the more recent version. Investigation suggested that perhaps half of the effects observed were attributable to the introduction of the mental arithmetic element mid-way through this period. The size of the overall effect would account for a significant proportion of the improvement in test results nationally over this period.
- Experimental comparisons suggested that KS3 Science test standards were similar in the 1996 and 2001 versions.
- The varied findings across key stages and subjects suggests that there has been no conspiracy to manipulate test results, otherwise all tests at all key stages would appear more lenient. It should be recognised that the task of those responsible for setting standards has been daunting given the pace of change as the new curriculum and testing regime has been introduced, refined and improved. Only now is the system settling.
- Investigations of children's perceptions of national tests via structured interviewing, in all subjects and key stages, showed them able to identify, and appreciate, salient features of the tests which have changed as national testing has evolved. Such developments largely aim to make the tests more accessible, more interesting and motivating and user-friendly; although the impact of such changes on performance is hard to gauge. The project's final report considers

various issues concerning such enhancements. It underlines the need to keep the effects of the renewal of curricular and assessment regimes in mind and to manage their impact so that they do not threaten the validity of assessments used for monitoring standards.

- The project team drew upon the project's evidence, and other experience, to make various policy recommendations designed to help maintain test standards. These include:
  - an integrated medium term cyclical approach to the management of curriculum renewal and national test development;
  - a strategy for test equating which replaces a year on year focus by a stepwise approach involving equivalence between successive tests and a 'stable' baseline measure, within each cycle of curriculum/assessment regime renewal, before moving – cautiously – to a new baseline when the cycle of curricular change necessitates;
  - a logical basis for prioritising evidence available when setting test thresholds: which pays due regard to national sample data and the inherent status quo, whilst simultaneously strengthening the search for sound (and transparent) evidence which should be seen as a pre-requisite for shifts in the pattern of results;
  - the longer term suggestion that teacher assessments (which the project's standardised test data suggested were relatively stable overall) might contribute to or determine national assessments for individual children, with (less intrusive and less costly) tests being used to monitor the system and direct any moderation of differences between schools and/or teachers.
- But the most important conclusion which can be reached from the project's work is tangential to our original brief. Taken alongside national test results over the period, the experimental evidence from all subjects and key stages indicated that there has been substantial real improvement in children's achievement. This was cross-validated by analyses of the LEA standardised test data made available to the project, wherever such data were relevant. The standardised test evidence, provided by several different sources, suggests that KS1 standardised reading scale scores improved by about 25% of a standard deviation, whilst in KS2 standardised reading scale scores improved by 10% to 16% of a standard deviation and standardised mathematics test scores improved by around 25% of a standard deviation. There have been significant gains in achievement in all subjects at all key stages, even those where our evidence suggests that national test results may be exaggerating their extent. Given that an extensive body of previous research demonstrates that system-wide improvements in achievement are generally small and hard won (Brooks, Foxman & Gorman, 1995), this should be seen as cause for congratulation to all concerned.

Impact

This project is likely to have a significant influence on research in this vein because of its methodological innovation. Not only is it the first large scale use of direct experimental comparisons, demonstrating their effectiveness over medium term time intervals of 3–5 years, but it is the first to collect allied analytic/empirical data to check on potential influences of curricular variations for the sample tested and any potential interactions with items contained in different test forms. This enabled systematic evaluation of the influence of such factors on the validity of the results. The project is also notable for the use of both replication and alternative methodologies to cross-validate comparisons. Confidence in research outcomes is much enhanced by replication, and when different approaches based on different data, assumptions, and definitions of equivalence, all point to the same conclusions – as was the case for this project, we can regard the findings as robust. The project's qualitative strand investigating pupils' capacity to identify and value enhancements to assessment instruments, and discussing how such issues relate to comparisons and the maintenance of standards in educational assessments, is also novel. More careful management of the introduction of enhancing features where comparisons over time are seen as a key use for assessments is likely to become a serious issue as awareness of their importance grows.

The project's substantive findings naturally attracted media attention. It made headlines even before the report's publication (e.g. 'Test result bombshell kept under wraps' – *Times Educational Supplement (TES)* 17.10.03) and attracted considerable news and editorial coverage on publication (e.g. 'Doubt cast on primary pupils progress' – *Guardian* 18.12.03).

It has continued to surface in news coverage from time to time since then and is now – two years on, in June 2005 – at the centre of a spat between the Statistics Commission (an official watchdog over the use of statistics by public bodies) and the Department for Education and Science (DfES). Professor Peter Tymms wrote to the Commission enclosing a paper (Tymms, 2004) – which extensively cites our conclusions – as evidence that national test statistics are unsuitable to represent trends in educational standards. The Commission's report (Statistics Commission, 2005) upheld key aspects of the complaint and suggested that official statements should indicate that improvements in KS2 test scores between 1995 and 2000 are in part due to factors other than rising achievement. (e.g.'QCA admits to 'illusory' primary test improvements' – *TES* 06.05.05). The QCA's evidence to the Statistics Commission had fulsomely supported the CoT Project's work and conclusions and, hence, the Commission's stance. However, the DfES objected to any restraint on the use of test results as evidence of the success of government policy and asked the Commission to think again: an approach the Commission has since largely rebuffed (e.g.'Commission stands by primary figures' – *TES* 03.06.05). So it seems unlikely that the matter has been laid to rest. Does it matter? Yes it does. Probity and confidence may be called into question if government appears to pick and choose between the convenient and inconvenient findings of well conducted research when assigning the credit and blame for past events. What of the implications for evidence-based policy making? For instance, might some of this project's conclusions have implications for the test targets which schools have been set for future years? Inappropriate targets are unlikely to serve their intended purpose. Policy making cannot ignore the facts.

Notwithstanding the brouhaha described above, the project's conclusions were in fact far from condemnatory about the management of the national testing system, which was in its infancy in the period concerned. Whilst problems were detected in some assessments, others successfully achieved the difficult task of maintaining test standards over a period of rapid change. The QCA and the (recently constituted) National Assessment Agency (NAA) and their contracted test development agencies have considerable technical expertise at their disposal and their good intentions are not disputed. But the project

(aided by the media coverage received – partial and inaccurate as it was at times) may well have a salutary effect on the thinking of those at the highest levels involved in policy making on assessment matters. Examining Bodies have long been aware that setting standards is a difficult and complex process; partly as a result of research into comparability issues. Putting national assessments under the spotlight will make the politicians and professionals managing them acutely conscious that the concept of error of measurement has real as well as theoretical aspects.

The Project suggested some quite fundamental improvements to arrangements for national tests. Some were implemented even before formal publication of the final report and it is understood that, partly in consequence of potential risks to the maintenance of standards having been highlighted by the CoT project, a wide-ranging review of the relevant features of the key stage test system is under way. Irrespective of whether our suggestions or alternative solutions are adopted, the project has served a valuable purpose in making policy makers aware of the need to treat the conceptual and technical aspects of educational standards with greater respect.

References

Bramley, T. (2005a). 'A rank-ordering method for equating tests by expert judgement', *Journal of Applied Measurement*, **6**, 2, 202–223.

Bramley, T. (2005b). 'Accessibility, easiness and standards', *Educational Research*, **47**, 2, 251–261.

Brooks, G., Foxman, D. & Gorman, T. (1995). *Standards in literacy and numeracy,* New Series 7. London: National Commission on Education Briefing.

Christie, T. and Forrest, G. (1981). *Standards at GCE A-level: 1963 and 1973*. London: Schools Council Publications/Macmillan Education.

Cresswell, M. (2000). 'The role of public examinations in defining and monitoring standards', *Proceedings of the British Academy*, **102**, 69–120.

Goldstein, H. (1983). 'Measuring changes in education over time: problems and possibilities', *Journal of Educational Measurement*, **20**, 4, 369–377.

Massey, A. (1994). 'Standards are slippery', *British Journal of Curriculum and Assessment*, **5**, 37–38.

Massey, A. and Elliott, G. (1996). *Aspects of writing in 16+ English examinations between 1980 and 1994*, Occasional Research Paper 1. Cambridge: UCLES.

Massey, A., Green, S., Dexter, T., and Hamnett , L. (2003). *Comparability of national tests over time: key stage test standards between 1996 and 2001*, final report to the QCA of the Comparability Over Time Project. Available from http://www.qca.org.uk/news/6301.html.

Newbould, C. and Massey, A. (1979). *Comparability using a common element*, Occasional Publication 7. Cambridge: TDRU .

Newton, P. (1997). 'Examining standards over time', *Research Papers in Education,* **12**, 3, 227–248.

Statistics Commission (2005). *Measuring Standards in English Primary Schools, Report No 23*, February 2005. London: Statistics Commission.

Tymms, P. (2004). 'Are standards rising in English primary schools?', *British Educational Research Journal*, **30**, 4, 479–493.

Accessibility, easiness and standards

Tom Bramley Principal Research Officer, Evaluation & Validation Unit

The following is a summary of a research article published in summer 2005 which was prompted by my experience of working in the National Curriculum test development group, formerly part of UCLES' Research & Evaluation Division, now part of OCR. One major task in the test development process is to carry out statistical and judgemental exercises which can provide evidence about where to set the cut-scores on the test. These cut-scores (level threshold boundaries, equivalent to grade boundaries on GCSEs and A-levels) are supposed to be at the same standard each year. Of course, tests can vary in difficulty and the cut-scores might not be at the same point on the raw mark scale each year, in order to allow for fluctuations in difficulty of the test from year to year (as with GCSEs and A-levels).

However, discussions about whether one year's test is easier or more difficult than the previous year's test can often get bogged down when the spectre of 'accessibility' raises its head. Is a 'more accessible' test the same as an 'easier' test? Are there any implications for where the cut-scores should be set if a test is deemed to be more accessible, as opposed to more easy? Is there any way to identify questions which are 'inaccessible'?

The main purpose of the article was to use a psychometric approach to attempt to answer these questions. The article begins by discussing the meaning of 'standards' and the ambiguity with which the term is used, particularly in media reporting of examination issues. The standard can be defined psychometrically as a point on the latent trait which is assumed to underlie or cause the responses to the test questions. The informal definition of statistical equating – that if standards have been correctly applied to two tests then it should be a matter of indifference to candidates whether they take test A or test B in terms of which level they obtain – is used as a starting point for discussing the issues raised by accessibility.

Three prototype arguments in favour of not raising the cut-scores by as many marks as the statistics might suggest when a test is deemed to be more 'accessible' were used to illustrate the discussion:

> *The paper is more accessible, but the amount of science hasn't changed.*
>
> *We've removed some of the hurdles which prevented the pupils from showing us what they can do.*
>
> *The pupils will be less 'turned off' by the paper and so we'd expect performance to improve.*