Statistics Commission

Measuring Standards in English Primary Schools

Statistics Commission Report No 23
February 2005

Statistics Commission

MEASURING STANDARDS IN ENGLISH PRIMARY SCHOOLS

Report by the Statistics Commission on an article by Peter Tymms

Introduction

1. This report looks at the use of official statistics of National Curriculum Key Stage test scores for monitoring standards in primary schools over time. Professor Peter Tymms of the Curriculum, Evaluation and Management (CEM) Centre at the University of Durham, wrote to the Statistics Commission on 26 March 2004, enclosing an article, 'Are standards rising in English primary schools?' that has subsequently appeared in the *British Educational Research Journal*[1], and inviting the Commission to consider the issues raised there. Professor Tymms' concern is that the statistics used – Key Stage 2 (KS2) test scores – are not suitable for the purpose of monitoring trends in standards over a period of years.

2. The Statistics Commission believed that the issues raised in Professor Tymms's paper were important in view of the high public profile of these statistics. Our approach to the investigation was:
    - to write to a small number of academics and educational organisations, enclosing the Tymms article and inviting their comments on it
    - to write to the Department for Education and Skills (DfES) and to the Qualifications and Curriculum Authority (QCA) with a similar request.

    We received three replies from six letters to academics and educational organisations. We also received a full response from the QCA. DfES said that policy responsibility for the issues raised lay with the QCA rather than with them, and therefore did not provide a reply.

[1] Tymms, P. (2004) *'*Are standards rising in English primary schools?*'* *British Educational Research Journal,* 30 (4)

Assessment of Tymms's article

Have standards risen?

3. Professor Tymms's article looks at the KS2 test scores for English and mathematics since the introduction of these tests in 1995. From 1995 to 2000, the aggregate KS2 scores showed a remarkable rise, with the percentage of pupils getting level 4 or above rising from 48 per cent in 1995 to 75 per cent in 2000 for English, and from 44 per cent in 1995 to 72 per cent in 2000 for mathematics. This rise in test scores has been widely publicised as evidence of a rapid rise in standards in primary schools. Since 2000, the upward trend in test scores has largely halted.

4. Tymms's article asks the question, 'are standards rising in English primary schools?' The evidence from the KS2 test scores, if taken at face value, suggests a very substantial rise in standards over the period from 1995 to 2000, but relatively little further improvement since then.

5. Tymms's comparison of KS2 scores over time with data from a range of independent tests and studies suggests that there was some rise in standards over the 1995 to 2000 period, but rather smaller than implied by the KS2 scores. The Massey report, commissioned by QCA to look at whether or not standards had risen, came to a similar conclusion. Tymms provides convincing reasons – in the incentives for teachers to teach test technique and to teach to the test – why introduction of a new 'high stakes' test, such as the KS tests, can be expected to lead to an initial rise in test scores, even if it does nothing to raise standards. He also provides a real world example – introduction of the Texas Assessment of Academic Skills – where this has happened in a similar situation.

View of academics

6. Overall there was support for the paper and its conclusions. In particular, there was general acceptance of the idea that standards had not risen as much as the KS2 test scores suggested.

View of the Qualifications Curriculum Authority

7. QCA described this paper as 'a useful synthesis of studies which yield data by which [National Curriculum] assessment data can be compared'. They nevertheless still believe that the findings of the study are less robust than the Massey report, which 'possessed a better research design in respect of triangulating National Assessment'.

This does not mean that Tymms's findings are wrong. As QCA note, 'the Tymms findings follow to some degree the Massey findings [although] they are not entirely in agreement'.

8. Regarding Tymms's comments on the experience in Texas in the 1990s following the introduction of new testing arrangements, QCA observe that 'the problems presented by the Texas test results … are a well-known phenomenon amongst measurement specialists. However, the implications … have – to date – not been well presented to government … nor have issues of measurement error and the implications … for performance tables'.

The Statistics Commission's view

9. **The Commission believes that it has been established that (a) the improvement in KS2 test scores between 1995 and 2000 substantially overstates the improvement in standards in English primary schools over that period, but (b) there was nevertheless some rise in standards.**

10. Ministers, and others who may want to use the test scores in a policy context, need to be made fully aware of any caveats about their interpretation. As Tymms's article demonstrates, the sharp rise in KS2 scores in the latter 1990s cannot be simply interpreted as a rise in schools performance standards – there are a number of qualifications that need to be made. Yet Government Departments have usually failed to mention any caveats about other possible reasons for rising test scores in their public comments. This may partly reflect a failure of communications on the part of the education research community.

11. **We feel that public presentation of the KS scores in statistical releases should include a clear statement about the uses to which the data may be put, and the limitations on the data in respect of those uses.** In that statement, it should be recognised that part of the rapid rise in test scores from 1995 to 2000 can be explained by factors other than rise in standards. We think it important that official statements should acknowledge the effects that teaching test technique and teaching to the test can have on test scores following the introduction of a new statutory test.

QCA procedures for maintaining standards

12. Tymms argues that one of the factors behind the sharp shift around 2000 in the trend in KS2 results from strongly rising to approximately flat was a tightening of the QCA procedures for setting cut-scores so as to maintain year-on-year consistency in test standards.

The Statistics Commission's view

13. The evidence that Tymms presents of a tightening of procedures by QCA is less than fully convincing, being almost entirely circumstantial and hinging on a switch from singular to plural in the description of the comparators used for the 'anchor test'. QCA have told us that they have been striving to improve their procedures over recent years, but that is not in itself evidence of the large step change in standard setting procedures that Tymms suggests took place.

14. The task of maintaining consistency of standards across time for the KS tests is a difficult one. The tests are high stakes and universal, so new tests need to be set every year. Another reason for setting a new set of tests every year is that the National Curriculum changes every year. Teaching to the test and teaching of test technique are probably inevitable in such circumstances; this will lead to some increase in average test scores. Finally, the process by which test scores are standardised – setting cut-scores that define boundaries for the 'levels' in which the results are presented – entails a loss of information that makes equating new tests with past tests more difficult.

15. Nevertheless **we are aware of no particular fault with the procedures that QCA now follow for maintaining test standards over time**. They are relatively complex – but this is not surprising given the various difficulties outlined above. A recent (December 2004) report by QCA's Independent Committee on Examination Standards[2] (which advises the QCA board, but is independent of QCA) comes to a similar conclusion – "[QCA] strategies for maintaining comparable examination standards over time do as well as possible". Although Tymms is critical of QCA procedures prior to 2000, the main problem he identifies appears to have been tackled and his article does not suggest that further improvements are urgently needed.

[2] McGraw, B., Gipps, C. and Godber, R. (2004) *Examination Standards.* Report of the independent committee to QCA.

Are statutory tests fit for the purpose of monitoring standards over time?

16. Tymms's position is that the substantial difficulties of maintaining a constant level of standards in a test like KS2 are so great that the test scores may just not be robust as an indicator of schools standards over time. His conclusion is that 'statutory tests must not be used to monitor standards over time'. They are simply not the right instrument. Test scores can and do move for reasons that have nothing to do with the standards that the children taking them have reached. Tymms's paper lists the characteristics of 'a perfect system for monitoring standards over time'. The statutory KS tests fail on most counts.

The view of academics

17. There was agreement that using the statutory tests to monitor standards over time presented substantial problems but there was no strong feeling that Tymms had found the solution.

The Statistics Commission's view

18. **As a theoretical proposition, Tymms is probably right. If the primary objective is to measure 'standards' in schools at an aggregate level over time, and the aim to design a system exclusively for that purpose, then the 'solution' may well be something like the 'perfect system' that Tymms describes** – involving 'the same secret tests used repeatedly on equivalent samples of pupils of the same age at the same time of year'.

19. The KS tests are clearly not a perfect system for monitoring standards over time at an aggregate level. However, this is not really surprising. The primary purpose of the KS tests is to measure the progress of individual pupils against the National Curriculum, not to measure 'aggregate standards'.

20. **It does not follow, however, that, because use of the KS tests is not the theoretically perfect way to measure aggregate standards over time, it is a wholly inadequate way of doing this.** Aggregate KS2 test scores gave a misleading picture in the early years of the KS tests, when test scores were improving rapidly at least partly for reasons other than rising standards. But it is not clear that they have given a misleading picture of trends in aggregate standards in the most recent years. The initial boost to scores from teaching test technique, etc, may now be played out. And the scope for

'drift' in standard setting procedures may have been reduced by QCA refining and improving their procedures for equating new tests with previous years.

Do we need an independent new body to monitor standards?

21. Tymms argues that we need a new independent body to monitor standards, along the lines of the old Assessment of Performance Unit (APU) of the Department of Education and Science, but independent of the DfES. His case rests on a number of propositions:
    - statutory test data are not fit for the purpose of monitoring aggregate standards over time
    - to monitor aggregate standards over time properly, independent data on standards are needed from a separate system of 'low-stakes' tests, designed for that purpose
    - administration of a separate system of tests, and analysis of data from it, should be entrusted to a new body (along the lines of the old APU) responsible for setting and measuring standards
    - that body needs to be seen to be independent of government.

The Statistics Commission's view

22. We agree that Tymms analysis has demonstrated that statutory test data are not ideal for monitoring standards over time. But that does not mean that the data are completely unsuitable for that purpose.

23. It should be remembered that the old APU was wound up in 1990. QCA have told us that the APU was not the stable, unproblematic mechanism often portrayed, and that it faced significant political and technical problems[3]. We think that it is necessary to understand why the old body, and the system it administered, were wound up, before a new system is launched.

Key Stage test data and performance targets

24. The key PSA targets for raising standards in schools are defined and measured in terms of the proportions achieving a specific level in the KS2 (primary) and KS3 (secondary) statutory tests. Tymms's article argues that statutory test data should not be used to monitor standards over time. This suggests that these key PSA targets are being evaluated by measures that are not fit for purpose.

[3] See Goldstein, H. and Gipps, C. (1083) *Monitoring children – a history of the APU.*

The Statistics Commission's view

25. KS test scores may not be an ideal measure of standards over time, but it does not follow that they are a completely unsuitable measure for a PSA target. There is no real alternative at present to using statutory test scores for setting targets for aggregate standards.

26. It is not clear that an aggregate measure from data from a new APU-type system would be a preferable measure for setting PSA targets. KS test scores have a number of advantages as a basis for setting PSA targets. Performance can be measured at a variety of levels (so can be used to set local targets); the data are reasonably timely; the concept measured – proportion of a school year group that attain a particular level in a national test – is easy to grasp.

Secretariat
Statistics Commission
8 February 2005

ANNEX A

Comments on Tymms's article from the Qualifications and Curriculum Authority

Letter from Tim Oates, Head of Research & Statistics at the QCA

Allen Ritchie
Head of Research & Investigation
Statistics Commission
10 Great George Street
London SW1P 3AE

Dear Allen,

Re Peter Tymms' report: 'Are standards rising in English primary schools?'.

Please find the QCA's response to Peter Tymms' report: 'Are standards rising in English primary schools?'. Thank you for this opportunity to feed into the Statistics Commission's consideration of this important area.

Mechanisms for monitoring and maintaining 'standards over time' remain worthy of further conceptual and technical discussion. QCA is of course responsible - through the codes of practice for GCSE and GCE, and the protocols for National Assessment - for overseeing the actions which are required each year to moderate the different demands which are placed on learners by different tests in different years. But at the same time as regulating the systems (in awarding bodies and in the NAA) which are used for this, we also continue to undertake research and review on the conceptual and practical problems which are presented by maintaining standards over time in curriculum-linked tests and examinations in a context where the curriculum demands are constantly changing. I do feel that further discussion with the Commission on our emerging findings in this area would be most helpful.

Focussing on Peter Tymms' work, we undertook critique of his earlier work and have attached this as a detailed technical annex. I also provide here our overview comments on his most recent paper. In summary we feel that whilst the paper is interesting, the findings of the study are less robust than the Massey Report. The Massey work possessed a better research design in respect of triangulating National Assessment, and although the Tymms' findings follow to some degree the Massey findings, they are not entirely in agreement. As I outline below, we place greater faith in the Massey methodology.

Peter Tymms' most recent paper: 'Are standards rising in English primary schools'
Peter Tymms originally contacted QCA in 2002 in order to examine the emerging discrepancy between data on English from PIPS and from NC testing. It was clear at that time that there were problems with the PIPS data and the associated analysis, in respect of: the quality and characteristics of the PIPS tests; the poor linkage between the PIPS test and the national curriculum; the small 'core' sample size in the PIPS system. These objections were analysed by QCA R&S team (annex #1) and were submitted as a paper to the QCA

Executive. It was felt that although the Tymms analysis should be treated as an important indicator, the method and instruments were insufficiently robust to be considered as an independent measurement of 'national standards'. By contrast, the later study commissioned by QCA and undertaken by Alf Massey (UCLES) possessed far more robust method and provided the closest we can currently get to an 'independent metric' by which to judge standards over time.

Leaving aside issues of whether there is technical merit and any intelligibility in 'judging standards over time' (in the context of a changing curriculum, underlying improvement in general intelligence (the Flynn Effect), etc), Massey remains the most important independent measure.

Tymms' most recent paper provides a useful synthesis of studies which yield data (of varying degrees of quality) by which NC assessment data can be compared. His analysis (using standardisation methods) shows that the PIPS data, in particular, mirrors broadly the findings of Massey. It does not, however, follow exactly the patterns of variation from NC assessment data which are presented by Massey. These discrepancies are not explored in adequate detail in the paper.

Conclusion – reactions to the Tymms analysis
We conclude that the PIPS analysis is an important contribution to a critical debate on the approaches used to determine standards in NC assessments. However, it does not provide the robust independent measure which was provided by the Massey Report. The Massey Report confirms that standards have risen but not necessarily to the extent suggested by NC assessment outcomes. Problematic discrepancies were identified in the results in some subjects (English in particular) prior to 2000.

Tymms' paper includes a reference to analysis of the introduction of revised test arrangements in Texas. The problems presented by the Texas test results were outlined by Dylan Williams (whilst at King's College – he is now at ETS in the US) in presentations during the late 90s and are a well-known phenomenon amongst measurement specialists. However, the implications of this study have – to date - not been well-presented to government, nor have issues of measurement error and the implications these have for performance tables.

Finally, Tymms' arguments for an APU-style measurement of standards over time are too simplistic. The APU model has attractions (low stakes, stability of tests over time), but the APU was not the stable, unproblematic mechanism which is often portrayed in current descriptions of its activities and achievements. Gipps' and Goldstein's analysis of the history of the APU demonstrates the significant political and technical problems.

Yours

Tim Oates
Head of Research & Statistics
01 11 04