Qualifications and Curriculum Authority

Level setting 2007

National curriculum assessments monitoring report

December 2007

QCA/07/3420

Contents

Executive summary

The Qualifications and Curriculum Authority (QCA) is the regulator of external qualifications and national curriculum assessments in England, and is committed to securing a fair deal for learners. The Regulation and Standards division of QCA regulates the national curriculum assessments produced by the National Assessment Agency (NAA), a subsidiary of QCA, against a regulatory framework and code of practice.

This report presents the findings of QCA's monitoring of the 2007 national curriculum level setting process. It considers the various level setting meetings monitored in 2007 and shows the following.

- The meetings were found to be broadly compliant with the *Code of practice.*[1] There was only one instance of non-compliance noted, and it was thought that this instance was unlikely to affect the security of the process.
- There is one serious issue, which relates to the ability of the level setting exercise data used at final level setting meetings to model accurately the final results data. While not affecting the security of the meetings this year or leading to non-compliance with the *Code of practice*, this issue could have an impact in subsequent years and as a result should be properly investigated and rectified by the NAA.
- General improvement was observed at all stages of the process this year, with a great deal of good practice noted by monitors.

There are further observations, detailed throughout this report, which the NAA should consider when undertaking any future improvements to the level setting process.

[1] *National curriculum assessments: Code of practice* (QCA/07/2828) was published in January 2007 and is available on the QCA website: www.qca.org.uk.

Introduction

Regulating the national curriculum assessments

As the regulator of England's examination and testing system it is the responsibility of the Qualifications and Curriculum Authority (QCA) to ensure that learners receive a fair deal, and that standards are secure and consistent over time. This responsibility relates to public examinations, but also to national curriculum assessments, where the National Assessment Agency (NAA), a subsidiary of QCA, is responsible for both the production and delivery of the assessments.

QCA is committed to ensuring that the same rigour is applied to the regulation of national curriculum assessments as is brought to bear on public examinations, and that the regulation of the assessments is proportionate, accountable, consistent, transparent and targeted. Particular attention is therefore given to those processes within the development cycle through which standards are maintained.

An overview of the level setting process

Level setting is one of the main processes through which standards in national curriculum assessments are maintained for all subjects at all key stages. The level setting process involves setting threshold marks for each level of performance in the current year's tests to maintain the established standard. In practical terms this means that where performance of a certain standard is awarded a level in a particular year's test, that same standard of performance would be awarded the same level in any other year. The level setting process involves three types of meeting and a variety of statistical and judgement-based evidence.

First, a set of threshold marks is produced by test development agencies through statistical equating methods. These threshold marks are presented to a panel of NAA staff at a series of 'draft level setting meetings'. The attendees at these meetings evaluate the evidence presented by the test development agency and agree draft level thresholds and the samples of scripts to be scrutinised (script scrutiny ranges), which inform the next stage of the process.

The next stage, 'script scrutiny meetings', involves the production of a set of threshold marks for each of the levels covered by a test. This is done by comparing performance in current scripts with the established standard found at the thresholds in archive scripts. The marking programme leader and other senior members of the marking team for each test work through the script scrutiny range to decide which mark best represents the threshold standard established in previous years.

The threshold marks produced at both of these meetings are presented in turn at the 'final level setting meetings' along with level setting exercise data. These data, established through a representative sample of pupils' results taken early in the marking process, allows the effect of setting the thresholds at different points to be measured against the overall performance in previous years. At each of the final level setting meetings for all three subjects a panel of senior markers, researchers, developers and NAA staff evaluates all the evidence and comes to a decision over the final level thresholds that should be presented to QCA's chief executive for approval.

Regulating the level setting process

The Regulation and Standards division's national curriculum assessments monitoring team observes meetings during the level setting process and monitors them against the *National curriculum assessments: Code of practice.*[2]

After the final level setting meetings the national curriculum assessments monitoring team provide QCA's chief executive with advice on the outcomes of the level setting process. This advice allows QCA's chief executive, and through him the general public, to be confident that appropriate procedures have been followed, that the results are robust and that standards are secure.

Monitoring the 2007 level setting process

In 2007 the national curriculum assessments monitoring team observed a selection of the three types of meetings that form the level setting process. These meetings were selected on the basis of risks identified during the 2006 level setting process and during the development of the 2007 tests. Observers did not become directly involved in any of the meetings, but did monitor them for compliance with the *Code of practice*. They completed a pre-agreed list of questions in order to collect evidence of compliance and to record other observations.

The main sections of this report detail the observations made during each of the three types of meeting. Each section sets out the number of meetings that were monitored and then moves on to detail those instances where issues relating to compliance with the *Code of practice* were recorded. The sections conclude with a summary of 'other observations' in relation to each type of meeting. These observations cover issues that

[2] *National curriculum assessments: Code of practice* (QCA/07/2828) was published in January 2007 and is available on the QCA website: www.qca.org.uk.

could affect the future security of thresholds if they are not addressed, as well as positive findings and improvements recorded during the monitoring.

This report finishes with a set of conclusions drawn from observation across the whole process and offers recommendations on areas that might require some further consideration for 2008. This section also considers what the implications of this year's monitoring work might be for the national curriculum assessments monitoring team's 2008 programme of work.

Section 1: Draft level setting meetings

Draft level setting meetings monitored in 2007

Key stage	Total number of meetings	Number monitored
Key stage 2	3	1
Key stage 3	3	3
Across both key stages	**6**	**4**

Compliance with the *Code of practice*

These meetings were monitored against section 10a ('Draft level setting') of the *Code of practice*. From the four meetings observed for monitoring purposes, the overall level of compliance was very high, with no significant breaches of the *Code of practice*.

However, paragraph 293 of the *Code of practice* states that draft level setting meetings 'must consider the thresholds and associated ranges in turn, starting with the target level for each key stage (and tier, where relevant), then progressing onto the higher levels before finishing with the lower levels'. Yet in one meeting (key stage 3 mathematics) the level thresholds were dealt with via a 'lowest to highest' approach that contradicted the requirement to start with the target level (level 5 in this instance). This is a minor issue and seemed unlikely to affect the security of the meeting's outcomes.

Other observations

There were noticeable improvements to the process in several of the meetings that were monitored this year. In particular, strengthening of several of the statistical equating exercises was evident. This was most prominent in key stage 2 science where actions initiated in previous years to improve the quality and consistency of equating evidence have now come into effect, with the meeting attendees now able to discuss and evaluate the results of several strong and broadly consistent equating methods.

This year, for the first time, all the statistical equating exercises were accompanied by a technical report that provided more specialist statistical information. In some areas the test development agencies went further than this by outlining the assumptions and limitations underlying their equating methods in their reports. This improved clarity and gave participants an increased sense of security in the decisions they were making.

This year also saw the removal of judgement-based exercises based around panels of teachers reviewing pre-test scripts. This meant that draft level setting meetings concentrated solely on statistical evidence. The NAA's decision to stop carrying out judgement-based work at the pre-test stage was carefully evaluated in the lead up to the draft level setting meetings this year and did not appear to have a detrimental impact on the ability to set draft thresholds or on the security of those thresholds.

The improved quality of the statistical equating exercises also had a positive effect on other stages of the process. There were several meetings that set very wide script scrutiny ranges in 2006, and this was highlighted as an issue that needed to be followed up in 2007, with several of this year's meetings chosen for monitoring as a direct result. This year the script scrutiny meetings were provided with more manageable ranges.

Section 2: Script scrutiny meetings

Script scrutiny meetings monitored in 2007

Key stage	Number of meetings	Number attended
Key stage 2	3	2
Key stage 3	4	1
Across both key stages	**7**	**3**

Compliance with the *Code of practice*

These meetings were monitored against section 10b ('Script scrutiny') of the *Code of practice*. Compliance was strong during the script scrutiny meetings, with no instances of non-compliance recorded at any of the three meetings that were monitored.

Other observations

As with draft level setting meetings, there were positive improvements in this year's script scrutiny meetings. This is certainly partially due to the average script scrutiny range being smaller than that of last year, but the praise that attendees gave to the test operations agency indicates an improvement in the general administrative processes surrounding the meetings.

For example, paragraph 300 of the *Code of practice* states that: 'The test operations agency, appointed by NAA to mark the tests, is also responsible for… ensuring that sufficient quantities of scripts covering the full range of marks are available for the meeting'.

In 2006 there were shortages of scripts for the script scrutiny meetings in a range of subjects. This year, however, this was not a problem. The number of scripts available was always in line with the requirements specified by the test operations agency, and there were often far more scripts available than these procedures required. The test operations agency also had additional scripts ready for the three marks above and below each script scrutiny range.

Some script scrutiny meetings set very narrow threshold zones of one or two marks, which were, on occasion, exactly the same as the final threshold recommendations. While this might occasionally occur in any key stage / subject, this year there were some key stages / subjects where it seemed to occur for every threshold. Scrutineers are asked

to establish both a zone and a threshold, suggesting that it is not a requirement to have threshold zones that exactly match the thresholds.

It was noticeable this year that NAA staff had a very clear understanding of their role as neutral observers of the script scrutiny meetings. They did not interfere or have any direct involvement in the meetings, even when one of the marking programme leaders invited the NAA to say if they were happy with a decision.

It is also of note that one of the meetings monitored this year was chaired by a new marking programme leader. The professional and assured way that the marking programme leader chaired this meeting demonstrated that a good continuity of procedures and processes had been achieved.

Section 3: Final level setting meetings

Final level setting meetings monitored in 2007

Key stage	Total number of meetings	Number monitored
Key stage 2	3	3
Key stage 3	3	3
Across both key stages	**6**	**6**

Compliance with the *Code of practice*

As in 2006, there were no issues of non-compliance observed during this year's final level setting meetings, which were monitored against section 10c ('Final level setting') of the *Code of practice*.

Other observations

Last year concerns were raised over the high number of participants at the final level setting meetings, but there was a very noticeable reduction in the number of participants at this year's meetings. Despite this, national curriculum assessments observers still raised some concerns about what the exact roles and responsibilities of some participants were, particularly when participants seemed to be advocating certain pieces of evidence. It often seemed, for example, that a representative from the NAA was present to support the test development agency without a similar advocate being present to support the marking programme leader.

All meetings clearly benefited from secure equating evidence and some members were strong in their support of this particular strand of evidence. In some subjects there are good historical reasons for favouring equating evidence. There are also other subjects where this is the first time for several years that equating evidence has been this secure, and where advocacy of script scrutiny would be more in line with past practice. In actuality, all of this year's script scrutiny and equating evidence was broadly comparable, so this conflict between script scrutiny and equating evidence did not become a significant issue.

While there were positive improvements in the script scrutiny and equating evidence, the third strand of evidence presented at meetings – the level setting exercise data – proved problematic (as a result of issues uncovered in previous years). It was apparent after the

level setting meetings in 2005 that there were significant differences between the level setting exercise and final data sets. These differences primarily seemed to be present in English and were most pronounced for key stage 3 writing.

There was a slight improvement in the results in 2006**,** but the final results for key stage 3 English in that year continued to be significantly different from the level setting exercise data, with the differences being similar to those in 2005. There were caveats attached to this year's data and these highlighted the issue for all attendees. However, the reasons for the discrepancies have not been confirmed as the NAA has yet to carry out research to investigate the problem. Until there is an investigation and the causes found, it seems unlikely that the problem can be properly addressed.

On the basis of the chosen thresholds for 2007, the level setting exercise data for key stage 3 shows a fall in the results for all three subjects, even taking into account the issues with the level setting exercise data. However, the selected thresholds do seem to be secure, given the alignment of the equating and script scrutiny evidence, and so it would seem that these thresholds maintain the established standard.

Conclusions and implications for future regulation

Compliance with the *Code of practice*

For the second year in a row there was an improvement in the level setting process when compared with the previous year, with a reduction in the number and severity of issues relating to compliance with the *Code of practice*. As the 2007 level setting process broadly complied with the *Code of practice* the results can be taken as secure and there can be confidence that standards have been maintained.

Issues requiring action by the National Assessment Agency

The issues relating to the level setting exercise data are severe enough to require a response from the NAA. This is particularly the case with the key stage 3 English data, which does not provide an accurate indicator of the final results.

The final results for 2007 have recently been published and a similar discrepancy for key stage 3 writing is once again present. Although this issue has been raised for the last two years, there has been no research conducted to discover the causes, which may have their origins in the live marking process. Until the causes of the problem are found, a caveat will need to be attached to the level setting exercise results.

It is recognised that the level setting exercise data play a different role at the meeting to that of either script scrutiny or equating evidence. Those two strands are used to maintain standards, while the level setting exercise data provides the meeting with an indication of the impact of potential thresholds. If the script scrutiny and equating evidence are in alignment, as happened this year, then a degree of security is brought to the threshold decisions, which makes the level setting exercise data almost irrelevant.

However, past experience indicates that these two strands cannot be relied upon to align every year, and while the level setting exercise data should not be treated in the same way as the script scrutiny or equating evidence, it should provide potentially useful information on the impact of the other types of evidence. Currently, the level setting exercise data cannot be relied upon to gauge the potential impact of the selected thresholds. This issue requires further investigation by the NAA.

Improvements and minor issues

There were strong signs of improvement across the process, and this was particularly the case with the draft level setting meetings, which all clearly followed a standard format. In those meetings where it was presented, information relating to the underlying

assumptions on which equating evidence was based helped to bring an increased level of transparency to the process, and ensured that secure and robust decisions were made. This approach should be carried forward to all of next year's meetings.

Script scrutiny also showed strong signs of improvement. NAA observers seemed very aware of their roles and the administration of the meetings was much stronger than that witnessed in the past few years. However, next year there is a transfer of responsibilities to a different test operations agency and this will need to be carefully managed by the NAA.

The final results from script scrutiny do seem to show an increasing precision, which is a perception that is not always supported by scrutineers' individual results. Threshold zones of two marks do not always seem adequately to account for the differences in scrutineers' judgements. For example, on one occasion a threshold zone was chosen which did not include a mark point at which six scrutineers had observed threshold performance. In such instances there does not seem to be a reason why such small zones have been identified and it might be appropriate if wider zones, which more accurately account for the judgements of all the scrutineers, were agreed.

The discussion and evaluation of evidence during final level setting meetings continues to be a strong aspect of the process. Some minor concerns were raised relating to roles and procedures at these meetings. However, these are not significant enough to affect the final outcomes and will be assuaged by a review of the NAA's document on formal procedures for the meetings.

Future regulation of the level setting process

Given this year's observations, the monitoring programme for 2007 was clearly appropriate and proportionate to current risk. As a result, and if everything remains the same, a similar monitoring programme would be suitable for the level setting process in 2008.

The monitoring of a representative sample of the draft level setting meetings will continue to be appropriate in 2008. As no particular 'at risk' areas have arisen during level setting this year, the draft level setting meetings to be monitored in 2008 will be identified on the basis of pre-test reports and any issues that arise before the meetings. One such issue is that different agencies are responsible for the development of the 2008 key stage 3 English and science tests from those who were responsible in 2007.

There will also be a new test operations agency responsible for parts of the process in 2008, and once again it will be important to ensure that standards are not affected by this

change of agency. The issue of ‘succession planning’ will require close scrutiny next year. This may have an impact on the numbers of script scrutiny meetings that are monitored in 2008. It may also lead to a review of the procedures that are in place to facilitate handovers.

Due to their high profile, all of the final level setting meetings will continue to be monitored in 2008, so that regulatory feedback can be provided to QCA’s chief executive. As a result, the national curriculum assessments monitoring team will need to be kept informed of the outcomes of those draft level setting and script scrutiny meetings that are not monitored, to ensure that the feedback given is accurate and fair.

The NAA will be contacted in due course to confirm the programme of work for 2008 and the meetings that will be monitored, to ensure that the process remains robust and that standards are secure.

