A Methodology for Aligning Assessment Tools to National Standards using Expert Judgements Informed by Student Performance

Introduction

National Standards in reading, writing and mathematics for years 1 to 8 have been established, and by 2011 all state schools are meant to produce overall teacher judgements against these standards for all students in these years. In order to support these judgements teachers are encouraged to use a range of assessment tools and processes, from standardised tests to learning discussions with students. One of the major issues for teachers will be knowing how to translate the outcomes of these tools into likely judgements against the appropriate standard. This problem of 'aligning' tools to the standards has similarities to the task of setting a minimum competency score on a test relative to a set of predefined mastery criteria, although there are important differences in the required outcomes.

Because of time pressures in developing information to support teachers in making judgements, it was important to find an appropriate methodology which was cost-effective, reasonably fast to implement, and made the maximum use of existing sources of data. The method chosen met these criteria, but is likely to be less robust than those based on extensive new data collection and analysis. For that reason the results produced should be regarded as provisional and subject to revision when more exhaustive data becomes available, from 2011 onwards.

In this paper we will:

1. Briefly review some of the existing literature on 'standard setting'
2. Describe the initial procedure which was developed for the 'tool alignment' exercise, and the rationale for it
3. Describe the results of an initial pilot exercise using this procedure, and the lessons learned
4. Discuss modifications to the procedure in the light of the pilot, and outline a protocol for a robust procedure for carrying out such an exercise in the future.

The main audience for this paper is academics and professionals who are interested in the methodology used to provide the initial assessment tool alignment information.

A brief review of some literature on standard setting

There is a large body of literature under the general heading of 'standard setting', which mainly deals with the task of determining a minimum score on a test which corresponds to some minimum competency standard, described verbally in terms of a set of criteria. The resulting 'cut-score' is intended to separate any group of candidates into those who, on balance, meet the minimum standards and those who do not. Stringer (2008) uses the term 'weak criterion referencing' to describe this process of setting cut-scores taking into account the difficulty of a particular test or exam.

The task we face in aligning assessment tools to National Standards is different in a number of ways from that described in the international 'standard setting' literature.

One of the main differences is that overall teacher judgements (OTJs) against the standards are to be informed by, but not directly derived from, test scores as part of a range of evidence about student performance. There is therefore no requirement to define a precise cut-score against a particular standard – in fact, extreme precision is unnecessary and undesirable, because we wish to allow for teacher judgement based on a range of evidence and therefore a more probabilistic mapping from test scores to standards is to be preferred. This consideration changes the purpose and focus of the procedure, and means that certain operations to increase consensus and improve precision are unnecessary.

However, our requirements and purposes are similar in a number of respects to those described in the 'standard setting' literature. The involvement of 'expert judges' who are experienced and knowledgeable in the given area, the focus on externally-defined criteria or standards, and the need to map numerical outcomes from assessment instruments on to the standards, are all similarities which lead us to believe we can learn from the existing literature. In addition, methods which involve consideration of records of authentic student performance (or 'scripts') are likely to provide insights which we can assimilate into the procedure we need to develop.

Kane (2002) gives a good overall summary of 'standard setting' methods, and distinguishes between 'test-centred' and 'examinee-centred' methods. In the former approach, judgements are made about the level of performance expected on different items or tasks in order to meet the standard at the lowest acceptable level. These judgements are combined to give the required cut-score. 'Examinee-centred' methods also use holistic judgements, but based on authentic examples of examinee performance to underpin them. This difference is quite crucial, because as Kane (2002) notes: "To the extent that the participants have experience in applying standards of practice, it is in applying them to the actual performance of practitioners in real practice situations, and the examinee-centered (sic) methods involve this kind of judgment". Näsström and Nyström (2008) mention "the difficulty for the judges to estimate the performance on individual items for a group of just qualified examinees" when discussing the Angoff test-centred method, but the difficulties of estimating the performance of hypothetical borderline individuals is common to all such methods. In general, therefore, for this kind of exercise judgements based on records of authentic student performance are to be preferred over those which focus on the attainment of hypothetical 'borderline' students.

Stahl (2008) reviews a number of standard setting approaches, including both test- and examinee-centred methods. One of the most commonly used methods in the former category is the Angoff method (see e.g. Ricker, 2003), in which expert judges are required to estimate the probability that a borderline successful candidate would pass each item in a test – simple aggregation of these estimates leads to the required cut-score. The method has the attraction of simplicity, but as Stahl (2008, p.6) notes: "the estimation of what percentage of the group of MCCs[1] will correctly answer a question is very difficult for SMEs[2]". Rickert (2003, p.28) states: "The ability of the

[1] Minimally competent candidates.
[2] Subject Matter Experts.

judges to conceptualize a minimally competent candidate is a problem that is difficult to overcome".

Another test-centred method is the Bookmark (see e.g. Schagen and Bradshaw, 2003). This uses actual performance data, mediated normally through an Item Response Theory (IRT) model, to arrange test items in ascending order of item difficulty. This removes from judges the task of determining relative item difficulty, and all they have to do is to read through the items and set a 'bookmark' after the last item they believe a minimally competent candidate would get right with a certain probability (usually set at 67%). Data from the different bookmarks set by different judges is then manipulated to set a cut-score. Although the number of judgements is essentially reduced to just one, as Stahl (2008, p.10) notes, judges often find this quite hard and there can be quite a wide spread in bookmark locations – although this is less of an issue when we are trying to capture the variability in judgements rather than set a single cut-score.

A complex set of analyses is commonly required to define the final outcome for the Bookmark, and the final outcome may not be intuitively related to the judges' decisions. Furthermore, to create the order of item difficulty requires extensive psychometric data on the items. Therefore, although the Bookmark can be a powerful method and makes use of performance data mediated by a psychometric model, its features and requirements make it more suitable for the analysis of larger-scale data collection and standards setting exercises (such as that based on NEMP 2010) than for the analysis of smaller collections of student performance records.

There are a number of other test-centred methods described in the literature. All have, or can be modified to have, input from actual performance data, sometimes in the form of feedback to judges on the consequences of their decisions on the distribution of results obtained for a sample of candidates to inform subsequent modifications to their judgements. However, in all these methods the focus is very much on judgements about the performance of hypothetical individuals such as the 'minimally competent candidate' or the 'borderline Level 4 student' and not on the results of real individuals. In examinee-centred methods, by contrast, the focus is very much on actual records of candidate achievement on the given test, and mapping these on to the externally-defined criteria.

Stahl (2008, pp.12ff) mentions a number of examinee-centred methods. The most holistic and qualitative is the 'body of work' method, which integrates samples of student work into the evaluation process - something along these lines may be suitable for more general research into overall teacher judgements (OTJs) based on multiple sources of evidence, and is clearly closely related to the process of moderating teacher judgements based on multiple sources of evidence (see below). The 'analytical judgement' method requires judges to assign samples of student work into performance categories, and then uses the lowest scores in one category in combination with the highest scores in the category below to define the cut-score. The 'paper selection method' provides judges with samples of student performance at each score point and requires them to select the sample which best represents the performance of a minimally competent candidate.

Hattie et al (2003) compare four different standard setting approaches for defining cut-scores in asTTle reading. They state (ibid, p.2): "It is not defensible to set up 'committees' to debate issues, decide on standards and then get some buy-in from other groups", a statement which chimes in with the general impression from the literature that it is better to use authentic exemplars of candidate performance to underpin this kind of work. The four methods they used were: 'modified Angoff', 'item signature', 'examinee-centred', and 'performance threshold'. The first two fall into the general test-centred category defined earlier, while the last two are generically examinee-centred. Comparison of the cut-scores produced by these four methods (ibid, p.23) showed a wide spread of results, indicating the difficulty in developing a robust process which delivers consistent results in this area.

Whetton, Twist and Sainsbury (2000) describe the process used in England to maintain consistent standards for national testing from year to year. Draft cut-scores are developed based on statistical equating of pre-test versions of one year's test with live scores from the same students on the previous year's test. These results are combined with test-centred judgemental methods (e.g. Angoff or Bookmark) to develop draft cut-scores in advance of the test being sat by all students in the year cohort. Before final cut-scores, and hence actual results, are published a sample of live test scripts is acquired and subjected to a 'script scrutiny' exercise. Judges are familiarised with scripts whose total score is at the cut-score set in previous years, and then sent packs of the current year's scripts and asked to identify scripts which are at an equivalent standard. Once a consensus is reached on this, the resulting total score is agreed to be the cut-score for the current year's test. Whetton et al (2000) state that this exercise "…has the advantage that the judges have real scripts from the actual test and hence from children with the correct levels of motivation and curriculum experience." They continue that "…such methods can work well if awarders share tacit standards, based upon guild knowledge, which are shared with the wider group of examination users."

This method has a number of points of similarity with the 'paper selection' method of Stahl (2008, pp.14f) and the 'examinee-centred' method of Hattie et al (2003), in terms of the selection of a range of authentic records of candidate performance and the requirement for these to be judged against predefined standards without the actual underlying scores being revealed to the judges. This class of methods has therefore been well-researched, and it seems that they contain the elements of a robust procedure which could be adapted to address the need to align assessment process outcomes to the National Standards in New Zealand.

Going forward, we have assumed the title 'script scrutiny' for our developing procedure, but with significant differences from the process described by Whetton et al (2000). The fundamental difference is in the purpose of the procedure: to capture a valid range of judgements against a numerical assessment scale, rather than to set a single cut-score which sharply divides the scale into 'passing' or 'failing' a given standard. For this reason the search for consensus and consistency between judges is not part of our process; on the contrary, we need judgements to be made and recorded independently in order to inform the measures of variability which are as important to us as any actual cut-score.

Initial design of script scrutiny procedure

In this section we will outline the methodology that was developed, based on the literature summarised above, to carry out script scrutiny exercises to collect data for informing the alignment of assessment tools and procedures to National Standards. The basic assumptions underpinning the initial design were:

- We have a valid assessment tool which is relevant to support teacher judgement against the given set of standards
- It produces outcomes which include a numerical score
- We have access to a number of existing records of authentic student performance on the tool ('scripts') which span a range of achievement sufficient to cover all likely judgements against the standard
- We can find suitably qualified judges who have experience of using the tool in classroom settings and are able to make judgements against the standards.

Given the above, the initial design was as follows.

We will divide the year level standards into two parts: years 1-3 and years 4-8, and do the exercise with separate teams of 10 judges for each assessment tool. For each year standard we will divide the assessment tool score into 10 score bands spanning the range of interest and collect 3 scripts for each such band. Each script will be judged by two different experts, giving 60 judgements per year standard per tool. Thus each judge will receive a pack of 6 scripts, randomly sorted and with total scores removed, on which to make judgements against the standard.

The allocation of scripts to judges needs to be made in such a way that each judge receives scripts covering a range of performance and as far as possible each script is given to a different pair of judges.

The overall activities to be carried out for each tool are set out below.

1. Determine a range of scores for each year standard which is likely to include the standard itself and a spread of performance either side.
2. Divide the score range for each standard into 10 score bands.
3. Locate 3 examples of student performance (scripts) within each score band.
4. Prepare copies of scripts, removing total scores and adding test questions (if not already on scripts) and any ancillary information (e.g. marking scheme).
5. Make up packs of scripts for the judges at each year standard.
6. Run the script scrutiny day with the judges (see below for details).
7. Collect data for the day and produce information to indicate mapping of scores to likely judgements against the standards.
8. Produce a short write-up of the methodology and results for general consumption.

A suggested structure for the script scrutiny day is set out below.

- A brief introduction setting the scene for the day and giving an overall briefing about what they are supposed to do.

- A session for each year standard (3 or 5 of these, depending on whether we are doing year 1-3 or year 4-8)
- A final session, in which we might show the results from the day and ask for comments and feedback on the outcomes of the process.

Within the sessions for each year standard, we would need to start with a brief introduction to the standard and a short discussion about the characteristics of a student who just barely met the standard. Then each judge would be given their pack of scripts and told to go and make independent judgements about the quality of each and where they would be likely to be placed against the standard: well below, below, at or above. They would record these judgements on a sheet which would be collected up before moving to the next year standard.

Notes on initial pilot exercise

An initial pilot exercise, based on the outline procedure set out above, was run at the Ministry of Education in Wellington on 20th November 2009. The focus of the exercise was the Observation Survey (see e.g. Clay, 1979), which is widely used in New Zealand to assess students' reading aptitude after one year at school[3]. The purpose of the exercise was to align the outcomes of this assessment tool with the National Standard for reading after one year[4]. 10 experienced judges (6 literacy professional development facilitators and 4 teachers) met together to examine batches of scripts which had been assembled, in order to assign each to one of the reporting bands: 'well below', 'below', 'at' or 'above' the standard. The procedure outlined above was followed in its essence, apart from the fact that just one standard was looked at in the course of the session, rather than 3 or 5. The judges were asked to fill in evaluation forms at the end of the session, which was a useful contribution to assessing the quality of the pilot exercise.

Two main occurrences should be noted. One was that the scripts to be selected only became available the day before, so it was not possible to carry out a proper quality control of these or to assign them to judges in a totally systematic fashion. The other was that during the session the judges requested a 'practice' script that they could judge collectively and discuss the results and their decision processes before judging the 'real' script. This request was acceded to, and the resulting discussion proved to be valuable to all concerned. Feedback forms were completed by all participants, and the following important comments were made:

- The shared discussion around the practice script was most useful, but the choice of this script heeds to be made carefully.
- It was important to share with the participants the generic formative definitions of the four reporting categories[5].
- The raw scores and stanine scores should be left on the scripts, and not blanked out, as these can be important elements in teachers' judgements.

[3] Also known as the '6 year net'.
[4] See http://nzcurriculum.tki.org.nz/National-Standards/Reading-and-writing-standards for details.
[5] See http://assessment.tki.org.nz/Effective-use-of-evidence/Overall-teacher-judgement-OTJ/A-student-s-achievement

- The quality of data recorded on the scripts was variable, with missing information or insufficient running records to make reliable judgements. There were a number of issues around the use of 'seen' and 'unseen' texts for running records, and this information was not always recorded.
- The ease with which participants made their judgements varied, depending on their previous experience with this assessment. Those with a Reading recovery background found it much more straightforward.
- Because the scripts were provided by some of the participants, there were issues about this, and this situation should be avoided in future.
- There was general agreement that the process was sound and valuable to the participants as a professional development exercise.
- The actual process of judging scripts was completed by all participants within about 30 minutes.

The data collected in the course of the exercise was fed back to the participants in simplified form, and then further analysis was carried out in order to provide guidance to schools on the range of possible judgements relative to both the average stanine score and the running record data. In the next section we will present a modified procedure for carrying out script scrutiny exercises in the future, based on the results of this pilot.

A modified procedure for script scrutiny

For each year standard and assessment tool, the total score should be sub-divided into 10 score bands spanning the range of interest and 3 scripts be chosen for each such band. Each script should be judged by two different participants (allowing us to estimate inter-rater reliability), giving 60 judgements per year standard per tool. Thus each participant would receive a pack of 6 scripts, randomly sorted and with total scores removed, on which to make judgements against the standard.

The overall activities to be carried out for each assessment tool to be aligned to each standard are set out below.

1. Determine a range of scores which is likely to include the standard itself and a spread of performance either side, including both 'well below' and 'above'. This will need to be based on the judgement of those familiar with the tool and the standard.
2. Divide the above score range into 10 score bands.
3. Collect a sufficiently large number of exemplars of student performance ('scripts') spanning the score range.
4. Quality assure the scripts, to ensure the relevant information is present and the scoring is accurate. Remove poor quality scripts.
5. Enter script identifiers and score values into a database.
6. Select (probably randomly) 3 scripts within each of the 10 score bands.
7. Prepare copies of scripts, removing student and school names and total scores, and adding test questions (if not already on scripts) and any ancillary information (e.g. marking scheme).
8. Make up packs of scripts for the judges at each year standard, with scripts randomised within packs (see below for how to allocate scripts to packs).

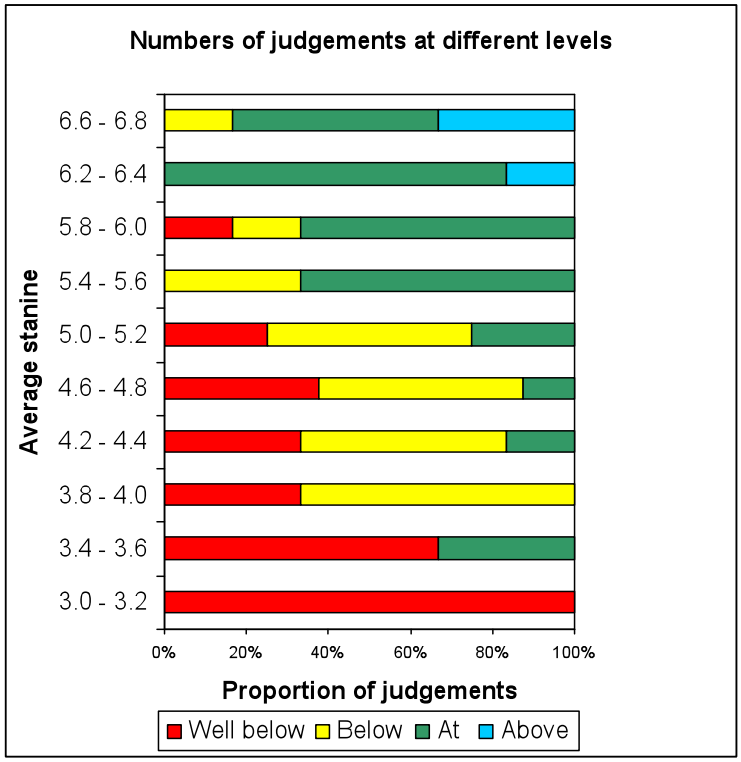
9. Run the script scrutiny day with the judges (see below for details).
10. Collect data for the day and produce information to indicate mapping of scores to likely judgements against the standards.
11. Produce a short write-up of the methodology and results for general consumption.

The allocation of scripts to participants needs to be made in such a way that each judge receives scripts covering a range of performance and as far as possible each script is given to a different pair of judges. An example of such an allocation is shown in Appendix A.

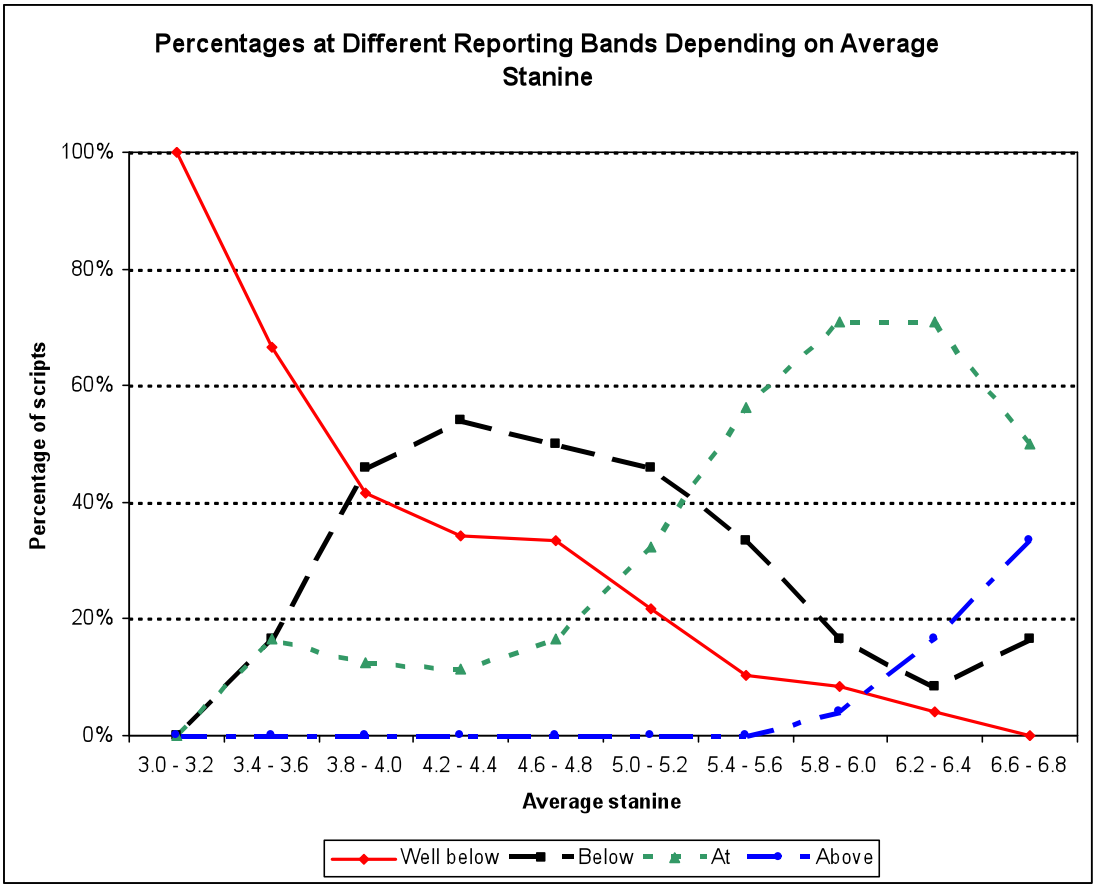
It is suggested that each script scrutiny day should focus on the standards for either years 1-3 or years 4-8. The overall structure for the day should have the following elements:

- A brief introduction setting the scene for the day and giving an overall briefing about what they are supposed to do.
- Presentation and discussion of participants' understanding of the four reporting bands for National Standards, based on the working definitions set out in Appendix A.
- A practice exercise with a common script, working in pairs, and with general discussion afterwards. If we are doing 5 standards (years 4-8) we may wish to have two practice scripts (e.g. years 5 and 7). The practice scripts should be carefully chosen to provide training in making judgements against the standards.
- A session for each year standard (3 or 5 of these, depending on whether we are doing 1-3 or 4-8). Details of these sessions are given below.
- A final session, in which preliminary results from the day are shown, and participants are asked for comments and feedback on the process and its outcomes, and fill in feedback forms.

The raw results for each year standard can be presented to the participants in the form of a simple graphic:



Further analysis of the data can produced 'smoothed' graphs illustrating the relationship between values on the underlying scale and the probabilities of being judged in each of the four bands.



It is important that the session facilitator emphasises initially that there are no expectations that certain numbers of scripts fall into the different judgement categories. If their professional judgement is that all scripts fall into 'well below', then so be it. It is also important to emphasise that participants need to make their judgements independently, so that the range of possible judgements for a given performance level is captured.

The relationship between script scrutiny and moderation

There is clearly much similarity between the process of script scrutiny as outlined above and the process of moderation recommended for developing overall teacher judgements against the national standards[6]. Both involve pulling together judgements based on different records of performance in order to form a coherent picture.

In script scrutiny, we are focused on a single assessment tool and multiple examples of performance against that tool. The aim is to bring together disparate and independent judgements in order to capture the possible variation in judgements that might be made based on a particular level of achievement as shown by this tool. The information collected has a universal purpose – to inform judgements against the standards nationally.

In moderation, we are focused on particular individuals with multiple assessment by different tools available, and wish to come to a shared understanding about the right judgement based on all the evidence. The information is specific to particular students, although the process is designed to guide judgements made on other students.

Although it is clear that participants in the script scrutiny exercises run by the Ministry have found these valuable for professional development, they should not be understood as a precise model for good practice in moderation.

References

Clay, M. (1979) *The Early Detection of Reading Difficulties: A Diagnostic Survey with Recovery Procedures.* Auckland: Heinemann.

Hattie, J.A., & Brown, G. T. L. (2003, August). *Standard setting for asTTle reading: A comparison of methods*. asTTle Technical Report #21, University of Auckland/Ministry of Education. Downloaded 16th November 2009 from: http://www.tki.org.nz/r/asttle/pdf/technical-reports/techreport21.pdf .

[6] See http://nzcurriculum.tki.org.nz/National-Standards/Key-information/Fact-sheets/Moderation

Kane, M. (2002) "Conducting examinee-centred standard-setting studies based on standards of practice", in *The Bar Examiner, Vol. 71, No. 4*, pp.6-13. http://www.ncbex.org/uploads/user_docrepos/710402_kane.pdf .

Näsström, G. and Nyström, P. (2008) "A comparison of two different methods for setting performance standards for a test with constructed-response items", in *Practical Assessment, Research & Evaluation, Vol. 13, No. 9,* pp.1-11. Downloaded 16th November 2009 from: http://pareonline.net/pdf/v13n9.pdf .

QCA (2007) *Level setting 2007: National curriculum assessments monitoring report.* Downloaded 16th November 2009 from: http://www.ofqual.gov.uk/files/QCA_Level_setting_report.pdf .

Ricker, K. (2003) *Setting Cut Scores: Critical Review of Angoff and Modified-Angoff Methods.* Alberta, Canada:Centre for Research in Applied Measurement and Evaluation, University of Alberta. Downloaded 19th November 2009 from: http://www.education.ualberta.ca/educ/psych/crame/files/RickerCSSE2003.pdf

Schagen, I. and Bradshaw, J. (2003) "Use of Bookmark for setting standards in reading tests', presented at the 29th IAEA conference, Manchester, October 2003. Downloaded 16th November 2009 from: http://www.emie.ac.uk/publications/other-publications/conference-papers/pdf_docs/UseofBookmark.PDF .

Stahl, J. (2008) 'Standard setting methodologies: strengths and weaknesses', presented at IAEA Conference, Cambridge, September 2008. Downloaded 13th November 2009 from: http://www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/180502_Stahl.pdf

Stringer, N. (2008) *An Appropriate Role for Professional Judgement in Maintaining Standards in English General Qualifications.* Guildford, Surrey: AQA. Downloaded 13th November 2009 from: http://www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/180490_Stringer.pdf .

Whetton, C., Twist, E. and Sainsbury, M. (2000) "National Tests and target setting: maintaining consistent standards", presented at AERA annual meeting, New Orleans, April 2000.

Appendix A

A possible scheme for allocating scripts to judges

Schematically:

Scores	Version A		Version B		Version C	
10	2	4	9	10	1	6
9	3	6	1	7	2	8
8	5	9	2	8	4	7
7	1	6	5	7	3	10
6	2	8	4	10	5	9
5	5	6	3	7	1	8
4	4	10	2	9	3	6
3	3	8	1	10	5	7
2	5	9	4	6	2	10
1	1	9	3	7	4	8

The red letters represent judges, and the versions A, B and C are the three different scripts within each score level. The above scheme would mean that each judge would receive packs with the following scripts:

1: 1A, 3B, 5C, 7A, 9B, 10C
2: 2C, 4B, 6A, 8B, 9C, 10A
3: 1B, 3A, 4C, 5B, 7C, 9A
4: 1C, 2B, 4A, 6B, 8C, 10A
5: 2A, 3C, 5A, 6C, 7B, 8A
6: 2B, 4C, 5A, 7A, 9A, 10C
7: 1B, 3C, 5B, 7B, 8C, 9B
8: 1C, 3A, 5C, 6A, 8B, 9C
9: 1A, 2A, 4B, 6C, 8A, 10B
10: 2C, 3B, 4A, 6B, 7C, 10B