An investigation into the test equating methods used during 2006, and the potential for strengthening their validity and reliability

Final report to the Qualifications and Curriculum Authority

Dr. Iasonas Lamprianou

University of Manchester and Cyprus Testing Service

September 2007

This report was commissioned by the Qualifications and Curriculum Authority (QCA).
The content of the report represents the views and conclusions of the author rather than those of QCA.

Contents

Executive summary

This research attempted to investigate the issue of the validity and reliability of equating methods used in national curriculum assessments to support standards over time. All of the Test Development Agencies (TDAs) gave responses which indicate a high degree of professionalism. According to their responses, the TDAs employ thorough and sophisticated methods to carry out equating tasks; however, the TDAs use very different methods to carry out very similar tasks. They offer reasonable, though not always well supported, arguments for why this is happening.

This research yielded a wealth of important findings, which are presented in this report. A few of the major findings are listed below. However, this list is not exhaustive: the reader is encouraged to go through the detailed comments in the next sections of this document as well.

Some of the main findings raised during this research are the following:

1. The TDAs use very different statistical models (i.e. Item Response Theory, equipercentile and linear equating) to carry out similar tasks:

    - A carefully designed research is certainly needed in order to investigate whether the 'competing' models give noticeably different results. One TDA argued that transition from one model to another needs to be done with care, because it may lead to different equating results. However, this is exactly the point: if different models give substantially (practically and statistically) different results, we need to know why we use the models we currently use. And we need to explain why we do not use other models.

    - In certain cases some TDAs declare that they may use Item Response Theory (IRT), if adequate evidence is presented to them that IRT is more efficient than their current techniques. In this case, QCA might wish to fund a relevant research to investigate the possible merits (or drawbacks) of IRT over other techniques in the context of the English National Curriculum tests.

    - Some of the responses of the TDAs gave the impression that they might choose to use IRT techniques if item-level data was available (implying that

item-level data is not always available). If this is the case, QCA may choose to help them acquire item-level data.

2. According to certain responses from the TDAs, the sample sizes are (at least in some cases) pre-defined by the NAA. However, some of the sample sizes reveal errors up to 1.5 marks in the mean of the scale (presumably much larger at the tails of the scale).

   - QCA needs to decide (and reason) why this error margin is acceptable. If an error of 1.5 or 2 marks at the mean of the scale (presumably much larger at the tails) is acceptable, we may want to investigate what % of students might have been awarded a different level, had the threshold been 1.5 or 2 marks up or down.

   - QCA may wish to explain why different error margins are currently acceptable for different subjects or different key stages (from the responses of the TDAs it is concluded that, in some cases, there are different equating errors for different subjects).

3. The assumptions of the statistical models in use may need to be investigated more thoroughly in some cases. If the assumptions do not empirically hold (practically speaking), and there is a poor model-data fit (for the intents and purposes of the test equating task), the test equating results may be misleading.

   - It has not been clarified in TDAs' responses if any of them ever faced similar problems in recent years.

   - If TDAs ever faced similar problems, it may be important to know what their reaction was.

   - It is not clear what the policy of the TDAs would be if such a problem ever appeared in the future. QCA may need to take a specific position on the issue and discuss this with the TDAs.

4. The test equating documentation could be enriched with additional information, in order to improve both transparency and reproducibility of results. This might facilitate/feed random and sporadic external audits to reassure the public about the quality of the procedures and the equating results.

- At the moment, the 'typical' documentation does not allow reproducibility/replication of the test equating procedures/results (i.e. it is not possible to reproduce the test equating results or experiment with the test equating data for academic/research purposes).
- As a result of the above point, random external audits (verifications or evaluations) are not possible. Random external audits once every few years might be a good idea in order to keep all parties involved in the test equating procedure alert. It could also reassure the public about the procedures followed and the equating results and provide justification for QCA to be more assertive and confident when supporting its level-setting procedures and results.
- Additional documentation, though, might come with a high time/money cost. QCA may need to find the correct balance between transparency and time (as well as fiscal) cost.

5. QCA might want to investigate ways to help TDAs publish their draft research findings without investing too much time setting up and formatting formal research papers for submission in journals or conferences. For example, an on-line library with technical reports, with adequate information and available datasets to replicate the analyses might be a very good research resource in order to guide decision making. It could also be a very good starting point for more research on technical issues that have to do with test equating methodology and statistical models. Most of the relevant research is now produced in other countries.
6. The investigation of the so-called pre-test effect is both academically and practically significant and needs to be investigated more systematically.
   - However, it is appreciated that it may be too costly and time consuming for a single TDA to undertake. Some coordination, and probably some joint research between TDAs might be more efficient, not only because economics of scale would save money, but because joint research could bring the researchers of the TDAs closer to adopting similar strategies and practices.
   - QCA may need to look into this research issue more closely. Some centrally coordinated (and probably funded research) would avoid replication of effort and could encourage all parties to jointly review/evaluate each others'

research. There are at least two strands of research: (a) to measure the pre-test effect, and if found to be important, (b) to find ways to overcome the pre-test effect for test equating purposes.

7. According to certain responses to the questionnaire, NAA specifies, where possible, the linking of new test results to those of the preceding year instead of to those of a baseline year. The TDAs, in their responses, do not rule out the possibility that this design might be susceptible to accumulation of error.

   - Further research on this would indicate whether accumulation of error is something TDAs should worry about.
   - If it is found that accumulation of error does occur, then it might be useful to investigate whether such an accumulation of error does not happen when the new test results are linked directly to the baseline year. If linking to a baseline year is less susceptible to accumulation of error QCA might consider changing the specification to link test results to the preceding year. It may also be useful to compare the methods on other characteristics as well: for example, large changes across years on the curriculum may make linking to the baseline year irrelevant.

Many of the above findings are not totally new, although they are stated more clearly and more directly in this document than they have been before. The Consultant was not surprised to find that many important issues had already been mentioned (though sometimes only indirectly) in previous reports and researches. For example, the accumulation of error when equating from one year to the next (instead of equating back to a baseline year), the pre-test effect, the need to justify why specific equating techniques are used (compared to other techniques), the need to be able to replicate/confirm the test equating analyses if necessary, etc, have been addressed indirectly or directly by the Massey *et al.* (2003) report.

It is important to mention here that the Consultant was not aware of the findings of the Massey *et al.* (2003) report before reaching his own conclusions. He deliberately chose not to read the Massey report because he wanted to check whether he would independently reach similar conclusions. Indeed, this external report reaches (independently) similar conclusions to those of the Massey *et al.* (2003) report.

Finally, it is very important to acknowledge that the TDAs did not give responses to all questions. There may be many explanations why this happened, for example a very long questionnaire; however, QCA might choose to raise some of the questions again (if QCA thinks that these questions are important).

Introduction

The background of the research

This report presents the results of a research commissioned to Dr. Iasonas Lamprianou by QCA in March 2006. Carrying out the research was deemed necessary by QCA in order to fulfil the principle of accountability (one of the five principles of good regulation) and in order to honour the requirements of the Regulatory Framework (validity and reliability: particularly with regard to the defensibility of the national curriculum assessments).

The aim and objectives of the report

The main aim of the research was to investigate the validity and the reliability of the equating methods used in national curriculum assessments to support standards over time.

The Consultant was commissioned to:

- undertake a survey of the relevant literature
- review the range of equating methods used for establishing threshold boundaries, and briefly comment on how those used in England compare with those used in similar contexts in other countries
- engage critically with the different equating methods used by the various TDAs and consider issues of validity and reliability
- write this report, following the QCA research report guidelines.

Methodology

The Consultant was commissioned to undertake the research after relevant discussions with QCA officers.

In order for the Consultant to be fully familiar with the context of the research, a number of relevant documents, research publications and other material were initially made available by QCA. An additional long list of potentially relevant publications and documents was also made available by QCA, so that the Consultant would be able to pick and study the most interesting and relevant documents.

The literature review proceeded throughout the duration of the research, and includes reports and documents published until the end of 2006, a few weeks before the final submission of this report.

In addition, the Consultant had the opportunity to participate in a relevant technical seminar where he met the psychometricians and statisticians of the TDAs, to become familiar with their work, the methods they use and to establish contact with them.

Following the seminar, and after studying the material presented by the TDAs during the seminar, a long and detailed questionnaire was constructed by the Consultant, and subsequently signed off by QCA.

The questionnaire was sent to the statistical experts at the TDAs in order to collect information about the methods and techniques they used to equate tests. The TDAs were given enough time to respond.

The responses of the TDAs staff were collected, studied in great detail, and clarifications were asked where necessary (and practical).

The final responses collected from the TDAs staff (including attached research papers and other notes) were studied and analysed in light of the continuous literature review, and the first results and initial comments of the Consultant were shared with QCA.

The Consultant was further asked by QCA to comment on the responses of the TDAs staff in writing, and to draw conclusions and suggestions which are presented at the end of this report.

The questionnaire

The questionnaire sent to staff at the TDAs was, along with the literature review, the major research instrument.

The questionnaire was constructed after a detailed study of the literature, and only after the Consultant felt that he understood the context of the research very well.

The questionnaire consisted of six sections each one focusing on one subject concerning the test equating tasks carried out by the TDAs:

1. Statistical models
2. Data-model fit, assumptions and properties of models
3. The quality of datasets/samples
4. Test equating design-test equating error
5. Software
6. Documentation

Each section included a short 'Description' sub-section, which described what the section was about, followed by a 'Statement' sub-section which presented the arguments and the philosophy of the 'Issues' raised in the next sub-section. Two of the sections closed with a sub-section called 'Food for thought' which could include ideas or opinions the TDAs could choose to comment on.

Although the questionnaire was long, it was hoped that the TDAs would find the time to respond in detail to all questions. Each of the sections was considered by the Consultant to be important enough in order to include in the questionnaire.

Methodological quality assurance procedures

The output of this research may be used to inform policy makers in England about important issues on the validity and reliability of test equating methods in the context of the English national curriculum tests. Therefore, it was very important for the Consultant to make sure that only valid, up-to-date and reliable information would be used in this report.

In order to ensure the high quality of the report, certain procedures were followed.

- Wherever possible, information based solely on internet sources was cross-checked by printed published material e.g. peer-reviewed academic journals. In contrast, information from the official web sites of large organisations or prestigious institutions did not need cross-validation with printed material.
- The intention of the Consultant was to contact all the institutions and organisations mentioned in this report in order to verify (cross-check) the validity of the material included, for example research reports, articles etc. Regrettably, this proved not to be practically possible in a few cases.

Ethical issues, copyrights etc

The initial (draft) report to QCA contained material that was confidential. Much of the confidential information was collected through QCA or was provided by colleagues on a confidential basis. In all cases it was necessary to reassure the sources of the information that the provided material would only be used for research purposes and to inform good practice.

Since this report may be in the public domain, the confidential material had to be removed, although this did not affect the results of the research, the conclusions or the findings of the report.

The format of the report

The report starts with the executive summary and the introduction which includes the background, the aims and objectives of the study and the methodology (including quality assurance procedures).

The literature review chapter presents a review of the methods currently used by the TDAs in England for test equating purposes, and also presents the basic concepts of test equating data collection design. This section draws on fresh literature but also on the Athanasou and Lamprianou (2002) book, cited in the references (the Consultant is the co-author of the book). It was a conscious choice on behalf of the Consultant not to present equations, or technical material, because this report is aimed for the informed public as well as the interested psychometricians and statisticians. However, because of the nature of the research, some of the literature was not practical to be presented here. It is better presented in the discussion section for each question and it helps the reader understand the comments of the Consultant to the responses of the TDAs.

The main body of the report includes the responses of the TDAs to the questionnaire and the comments of the Consultant on their responses. It is important to note that the views expressed in these responses are those of experts who were involved in the test development process at the various TDAs during the 2006 development cycle. They should in no way be seen as the official policy or opinions of the organisations themselves. Most sections include references to the literature that may not be presented in the literature review.

The report ends with the discussion and recommendations, and references.

Literature Review

Test equating

Multiple forms of the same test, built to the same content and statistical properties, but containing different (or at least some number of different) items, are frequently used for security purposes and to compare changes in performance across time. The use of different forms of the same test (or different tests aiming to measure the same constructs from year to year, as happens in the UK), raises the issue of the comparability of test scores. Although many resources may be spent in order to build multiple forms of the same test that will be parallel in structure, timing, item types, format and subject matter, the actual test difficulty will inevitably vary. Therefore, it is not possible to claim that the test scores on different forms are directly comparable, but a more formal 'test equating' is needed.

After two tests are equated, pairs of equivalent scores become available. For example, such a pair of equivalent scores could be (13, 15) which indicates that a total score of 13 on the first paper is equivalent to a total score of 15 on the second paper. To keep things simple, in order to compare achievement using two different tests, one simply needs to use a conversion table or graph to convert the scores of one test to the equivalent on the other test.

The widespread use of high stakes public examinations, and the pressure on the psychometricians to be able to interpret results from administrations of different tests, have generated an increased interest in the area of test equating research and development. For example, the work of Kolen and Brennan (2004) was a result of the increasing need of the academic community (and the practitioners as well) for a comprehensive book about the fundamentals of test equating (and related concepts). Kolen and Brennan followed the tradition of Angoff (1971), Holland and Rubin (1982) and others in giving detailed and elaborated accounts of the most prominent test equating methods, their advantages and their drawbacks.

A possible definition of test equating could be the one proposed by Kolen and Brennan: '*equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably*' (p.2). Note that in the English context we do not talk about different forms of tests, but about different tests,

assumed to be measuring the same construct – a construct that is defined by the national curriculum.

In order to achieve the goal mentioned above (actually, a definition of equating), many different methods have been proposed, the most prominent of which may be grouped as Classical Test Theory (CTT) approaches, or Item Response Theory (IRT) approaches. The usual CTT approaches used in England are the linear and the equipercentile methods, whereas the most usual IRT approaches are the 2-Parameter Logistic Model (2PLM) and the Rasch Model (RM).

A detailed technical description of the methods, however, is beyond the scope of this research which aims to keep a rather non-technical style, avoiding equations and complicated figures. However, for the sake of the interested reader, some 'popularised' details will be given on the data collection designs for test equating. This is necessary in order to put the discussion that follows, (as well as the results of the research) into a relevant context.

Data collection designs for test equating

Some authors have extensively discussed several designs of data collection and the test equating methods resulting from them. One of the simplest data collection designs is the 'Single-Group Design'. According to this design, two forms of the same test are administered to the same sample of students. The following table exhibits the data collection design for two papers.

	Test	
Sample	Paper 1	Paper 2
Common Sample	√	√

Table 1. The Single-Group Design for test equating

Using this design means that we are ready to accept that the examinees' score on the second test is not affected by the experience gained from taking the first test some time ago. It is also assumed that factors like learning, practice or fatigue do not, at least practically, affect the test results on the second test.

However, a more complicated data collection design for the equating of tests can take the form of the 'Anchor-Test-Nonequivalent-Groups Design'. According to this design, each of two different groups of examinees completes a different form of the same test. Still, both groups complete a common test, say V, which is called 'the anchor test' and is used as a link between the two forms. It is crucial though that the anchor test should be very similar in content and difficulty to the tests to be equated. The following table exhibits the Anchor-Test-Nonequivalent-Groups Design, which can be employed for the equation of two papers through an anchor paper.

	Test		
Sample	Paper A	Paper B	Anchor Paper
Sub-sample 1	√		√
Sub-sample 2		√	√

Table 2. The Anchor-Test-Nonequivalent-Groups Design for test equating

The term 'anchor test' describes a set of items, which can be used as a link for the equating of two or more tests. According to Petersen *et al.* (1989), the anchor test should be administered to both groups in the same order, bearing in mind that scores on a second test are always affected by the test previously administered (learning, practice, fatigue and so on). Finally, an anchor test can either be 'external' or 'internal'. If the anchor test is just a sub-test in the tests to be equated then it is called an internal anchor test. If, however, an anchor test is a different form tested at a different time, that test is an external anchor test.

Single-group test equating

Let us consider the case where two mathematics papers, say, 1 and 2, were completed by the same students. If the students completed the two tests in a certain order, for example the schools always administered paper 1 first, then issues such as fatigue or learning may arise.

Let us assume that Figure 1 illustrates the distribution of total scores on the two papers. Two papers (paper 1 and paper 2), very similar in content, item types and timing were administered to the same group of pupils. The vertical axis of Figure 1 indicates the number of students that achieved each of the scores of the horizontal axis.

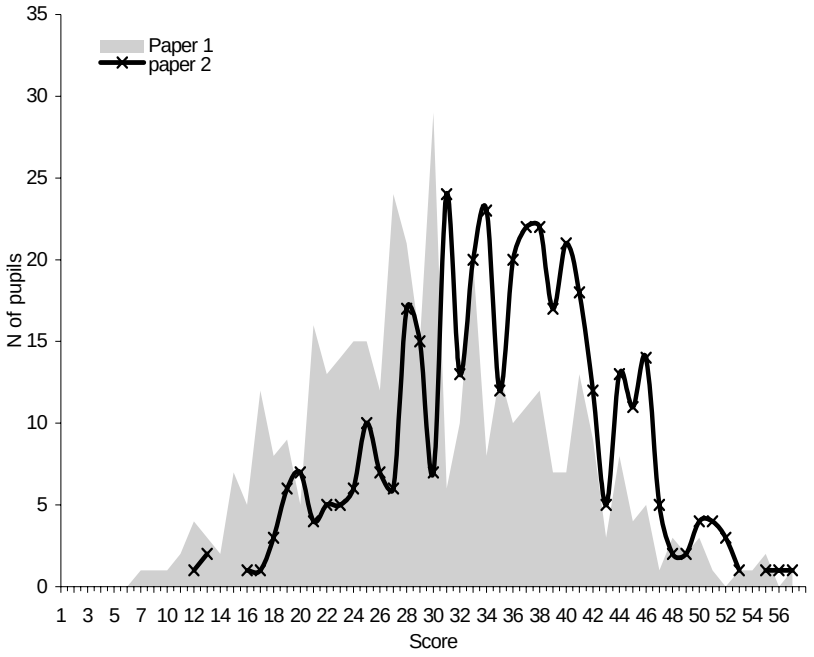

Figure 1. The distribution of test scores on papers 1 and 2

The scores on paper 2 are generally larger than the scores on paper 1 and therefore, paper 2 is easier than paper 1. However, the two distributions have reasonably similar shapes and this is encouraging for the use of the equipercentile and the linear test equating method.

The linear equating procedure (for groups of different ability) was described by Angoff in 1982 as well as by Levine in 1955. This method may be used either for the equation of equally reliable, or unequally reliable tests, with some modifications. In order to employ this method of test equating, some assumptions are made. First, the tests (forms) are assumed to be parallel. Levine (1955) assumes that all tests to be equated should be parallel not only in item types, format and timing but in structure, and subject matter, allowing though for different degrees of difficulty.

The equipercentile equating uses the cumulative percentages in order to equate two tests. Linear equating is considered to be a simplified equipercentile equating (this will be discussed later), so this section will mostly focus on the equipercentile equating.

The cumulative percentages (in other words the percentile values) are values of the total score that divide the data into two groups so that a certain percentage of the

sample is above and the rest of the sample is below. For example, the 95th percentile indicates the value of the total score below which 95% of the students fall. Table 3 indicates the distinction between percentages and cumulative percentages.

Total score	Frequency	Percentage	Cumulative %
6	0	0	0
7	1	0.25	0.25
9	1	0.25	0.50
10	1	0.25	0.75
11	2	0.50	1.25
12	4	1.00	2.25
13	3	0.75	3.00
14	2	0.50	3.50
15	7	1.75	5.25
….. (table continues) ….			

Table 3. Percentages and cumulative percentages

Table 3 indicates that the cumulative percentage for a score on the test is the sum of the percentage of the students who obtained this score and of all the previous percentages. In other words, it is the percentage of the students who obtained this score or obtained a lower score. Figure 2 illustrates the cumulative distribution as described by table 3.

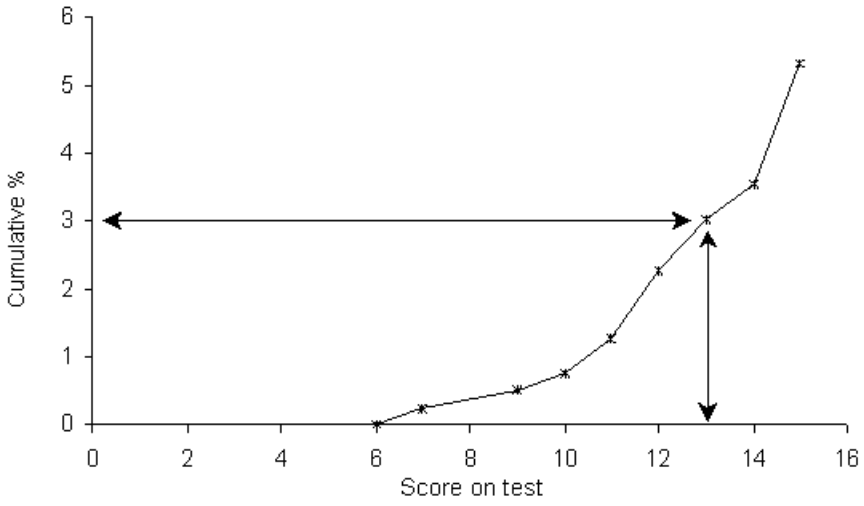

Figure 2. The cumulative distribution of scores

You can see from Figure 2 that by drawing horizontal and vertical lines we can find the scores that correspond to a specific percentile. For example, we can find that the percentage of students whose score is equal or below 13 marks is 3%.

The percentages and cumulative percentages are the only ‘statistics’ you need to know in order to employ the equipercentile equation. Now that we have clarified this issue,

we can declare the aims of the equipercentile equating method. The aim of the method is to make the raw scores on two tests correspond to the same cumulative percentage for a certain group of examinees. The technique demands either the conversion of one set of scores to the other, or the conversion of the two sets of scores to a third (new) one. In other words, after successful equating, it should be indifferent to the examinees whether they have sat one paper or the other.

Therefore, for the equipercentile equating of two tests, it is enough to use the graphs of the cumulative percentages of their scores. If the two graphs are drawn on the same axes, then the raw scores on each test that correspond to the same cumulative percentage can be identified. These scores form pairs of equivalent scores. Many such pairs can be plotted on a graph to form the 'conversion line' which is the function that transforms the scores of the one test onto the score scale of the other test.

A major advantage of the equipercentile equating between two tests X and Y is the opportunity to set cutting scores (for groups similar to that used for the equating) on both tests and still be sure that the same percentage of examinees will succeed at each test. As mentioned by Kolen, no other equating technique possesses this method. Figure 3 exhibits the cumulative test score distributions for the two papers 1 and 2.

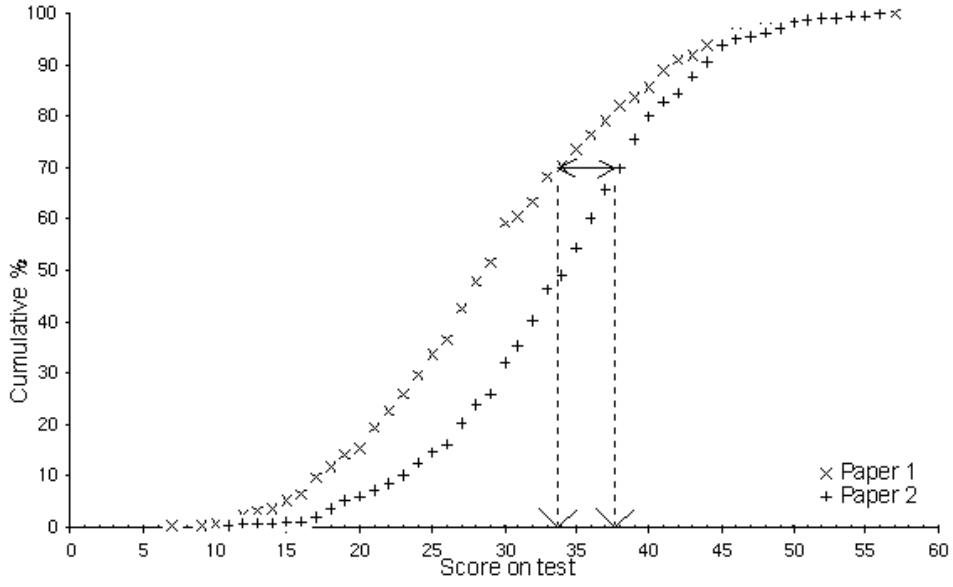

Figure 3. A pair of equivalent scores between papers 1 and 2

The equipercentile equating method suggests that the score scales for two tests for a certain group of examinees are comparable if the score distributions have identical

shape. The two distributions on Figure 3 have roughly similar shape and thus a table of pairs of raw scores with identical percentile ranks may be constructed. For example, the two vertical arrows on the graph show that 70% of the examinees have a score smaller or equal to 34 on paper 1. The same cumulative percentage on paper 2 indicates that 70% of the students have a score which is equal or smaller than 38. Thus, a pair of equivalent scores may be identified as (34, 38) since the two scores share the same cumulative percentage on the two papers.

For the creation of the conversion line, all the possible pairs of equivalent scores should be identified. It is, sometimes, necessary to connect the consecutive points for each test with straight lines so that the equivalent scores between the two tests will be identified easily. In such a case, the cumulative distributions will look like Figure 4.

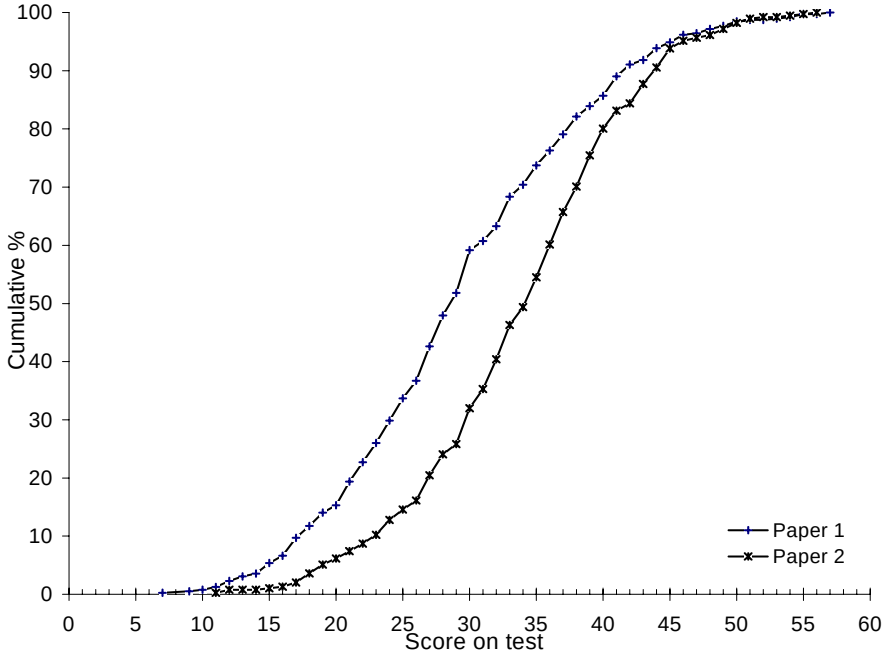

Figure 4. The linearly interpolated relative cumulative frequency distributions of papers 1 and 2

The two distributions on Figure 4 are the 'linearly interpolated relative cumulative frequency distributions' described by Petersen *et al.* The two cumulative distributions are plotted on the same axis. The distributions are called 'linearly interpolated' because the points consisting each one are connected with a straight line. The linear interpolation converts the discrete distribution of the Figure 3 to the 'continuous' distribution of Figure 4. That way, it is easier to find as many pairs of equivalent scores as desired.

For each cumulative percentage, (y-value), the corresponding test score, (x-value), is found for both distributions. For example, for the cumulative percentage 60%, the

scores for paper 1 and paper 2 are approximately 30 and 36 respectively. Thus, a pair of equivalent scores has been prepared (30, 36). Using this technique, the conversion line of Figure 5 was generated.

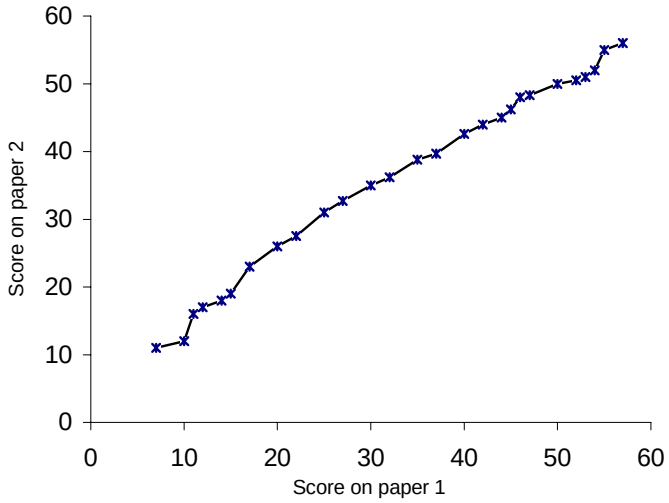

Figure 5. Conversion line between papers 1 and 2

Because of the irregular behaviour of the conversion line, especially at the edges, smoothing techniques may be used. Various analytic smoothing techniques have been proposed and used. With smoothing, the line produced need not pass through all the observed points. Two smoothing techniques are shown at Figures 6 and 7.

The smoothing of Figure 6 was achieved by applying a straight line on the conversion line of Figure 5. Although the eye may give the impression that the straight line loses much of the information of the graph, the R^2 statistic, (a statistic used to indicate how well a line describes a dataset), has a value of 0.99 which means that the straight line explains 99% of the variance of our data.

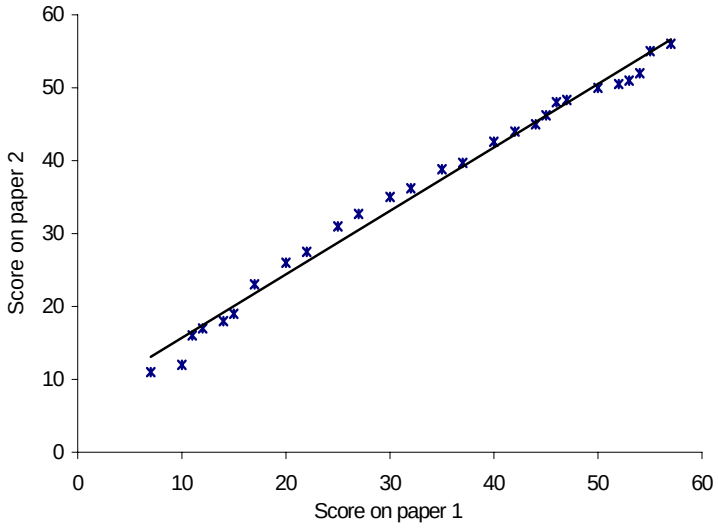

Figure 6. Linearly smoothed conversion line between papers 1 and 2

Figure 7 indicates a non-linear smoothing. This line seems to have an even better fit but the improvement over the previous smoothing is almost negligible and does not deserve the extra complexity of a non-linear smoothing. The R^2 in Figure 7 approaches 1 (almost all variance is explained).

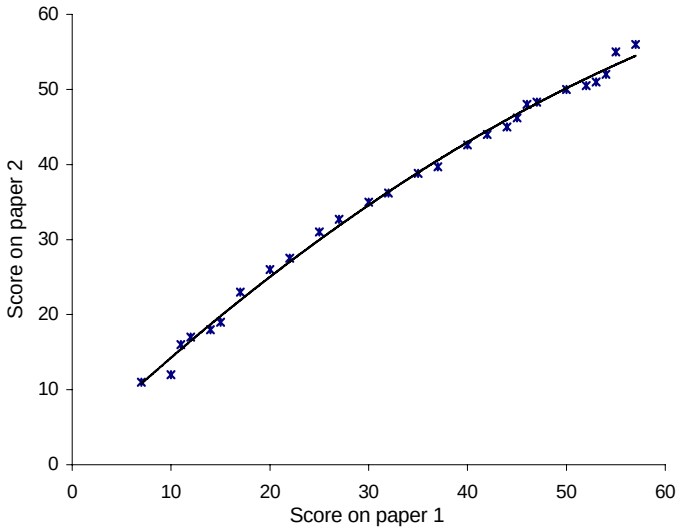

Figure 7. Non-linearly smoothed conversion line between papers 1 and 2

One of the concepts that may confuse readers is smoothing. Smoothing may be applied before the test equating (called pre-smoothing), or after (post-smoothing). Pre-smoothing, as well as post-smoothing, techniques may be used/benefit a test equating task (e.g. Kolen, 1984, 1991; Livingston, 1993; Hanso *et al.* 1994) by smoothing out any irregularities, spikes or zigzags (see the figures above) of the score distributions caused by many factors like small sample size, atypical sampling etc. Recently, Kolen and Brennan (2004) provided a long list of pre- and post- smoothing techniques. Smoothing might be useful in the context of the English TDAs but again, there is no clear evidence either supporting or opposing this argument. The Consultant decided to include a question concerning smoothing in the questionnaire in order to capture the opinion of the TDAs on this issue.

Smoothing is not a panacea and should always be treated with much caution. The symmetry of the conversion line can be easily destroyed. Moreover, research has indicated that smoothing techniques may introduce bias even for large samples in some instances. For this reason, smoothing should always be handled with care.

Anchor-Test-Non equivalent-Groups equating

Under this design, group A is administered form X and group B is administered form Y. Both groups also complete a common form V. Using the equipercentile method already explained above, forms X and V are equated using data from group A. Then, using data from group B, tests Y and V are equated. For each score on form V it is now possible to locate the scores on X and Y. The distribution of the pairs of scores (x, y) are plotted, and maybe smoothed, to form the conversion line between the tests X and Y.

The rest of this section will provide a practical example. Two different groups of pupils were administered three tests. Group A was administered paper 1 and paper 3 and group B was administered paper 2 and paper 3. Paper 3 will be used as the anchor paper since it was administered to all the pupils.

The sample of pupils that completed the anchor test (paper 3) may be divided into two sub-samples. The first sub-sample is the group of students who completed paper 1 and the second sub-sample is the group of students who completed paper 2. Following the process of equipercentile equating already described, pairs of equivalent scores were defined between the anchor test and paper 1. Figure 8 is the conversion line between the anchor test and paper 1.

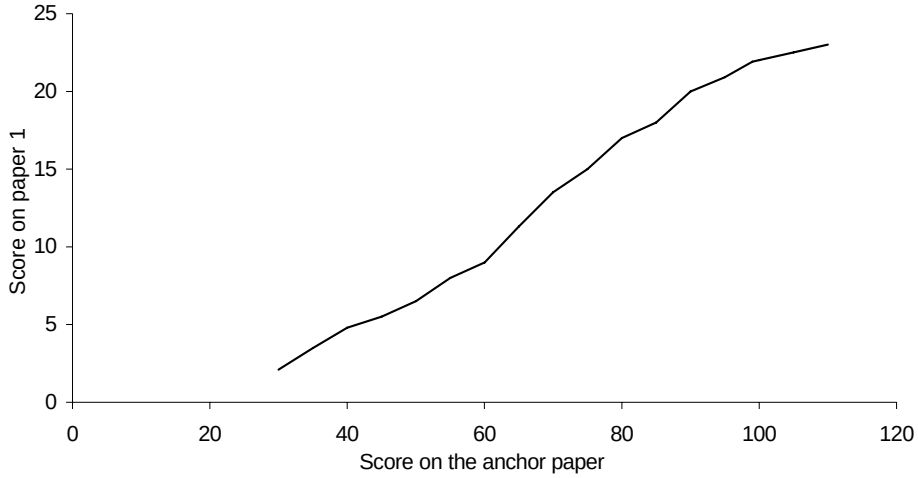

Figure 8. Conversion line between paper 1 and the anchor paper

The same process was followed in order to equate the anchor test to paper 2. Figure 9 illustrates the conversion line between paper 2 and the anchor test.

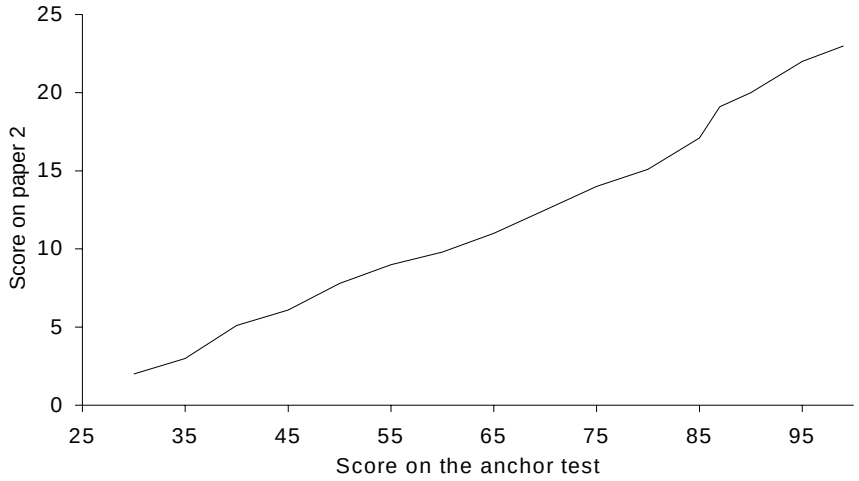

Figure 9. Conversion line between paper 2 and the anchor paper

The two previous conversion lines will now be used for the identification of the equivalent scores between paper 1 and paper 2. For every score on the anchor paper, a corresponding value on both papers 1 and 2 can be defined. Thus, having the anchor paper as a link, pairs of equivalent scores on papers 1 and 2 can be defined. Using this technique, Figure 10 demonstrates the conversion line between paper 1 and paper 2.

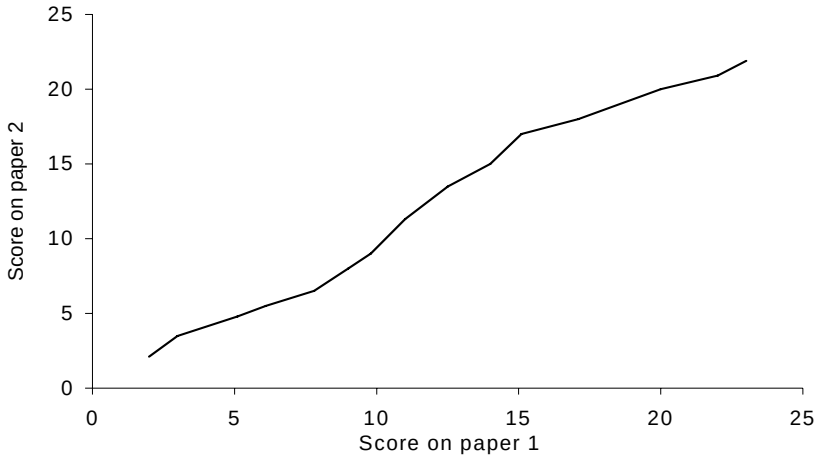

Figure 10. Conversion line between paper 1 and paper 2

Test equating using Item Response Theory

The first thing that comes to mind is 'why do we need the IRT for test equating?' Well, the previous test equating attempts using CTT cannot easily overcome the equity, symmetry and invariance problems. However, this is not the case for the IRT, provided the model fits the data (Kolen, 1981). These problems (equity, symmetry and invariance) are the basic properties of IRT models (Hambleton and Swaminathan, 1985; Lord, 1980).

Moreover, Kolen argues that both equipercentile and linear equating techniques under CTT may be unsatisfactory when the tests to be equated differ substantially in difficulty, while the use of IRT, according to some researchers, may totally discard the need for test equating.

The classical approach of analysing test responses is heavily dependent on the sample that completed the test. For example, item easiness is analogous to the percentage of the examinees that got the item correct. Item quality is usually estimated by the point-biserial correlation between item-response and test total score. In addition to that, a person's ability is usually defined by the percentile he/she is standing in the examinees' population. Kolen (1981) clearly stated that '*…conventional equipercentile or linear equating can be strictly used only with parallel tests…In theory, Item Response Theory models… can be used to equate both parallel and non-parallel tests…*' (p.1-2). So IRT is not only 'sample-free' but 'item-free' as well (at least, this is what the theory claims when the data fit the model perfectly well).

The degree of dependency of test calibration on the abilities of the examinees for each of the two methods (CTT and IRT) can be checked by applying an experiment described by Wright (1967). According to the supporters of IRT, since the CTT test equating methods are based on traditional statistics which vary according to the sample used, these methods will be severely affected by the sample of the examinees' used. The supporters of IRT also claim that whatever the sample (provided the model holds), the test calibration is person-free; therefore, IRT has a serious advantage over the CTT methods on this aspect.

Since the invariance of item parameters is assumed by the model, there is no reason for test equating. The item parameters, once calibrated, can theoretically be used to estimate the abilities of any group of examinees. However, it is often the case that differences on the item difficulties may arise when calibrated with different groups. Hambleton and Swaminathan (1985) have explained that such a case occurs because of the arbitrary fixing of the ability metric at zero during the items' calibration. Since there is always a linear relationship between the item parameters and the ability parameters, if we fix the mean value of the ability θ, this will affect items' difficulty and vice versa.

So, in an IRT test equating situation, there will be a linear relationship between the estimated abilities on each test (Hambleton and Swaminathan, 1985, p.203-204) expressed as:

$$è_{paper1} - ì_{è\ paper1} = è_{paper2} - ì_{è\ paper2} \quad \Rightarrow \quad è_{paper1} = è_{paper2} + (ì_{è\ paper1} - ì_{è\ paper2})$$

(for the Rasch model and the mathematically equivalent 1-Parameter Logistic Model)

The respective equation for the 2- and 3-parameter logistic models is given below:

$$è_{paper1} = (ó_{è\ paper1} / ó_{è\ paper2})\ è_{paper2} + [ì_{è\ paper1} - (ó_{è\ paper1} / ó_{è\ paper2})\ ì_{è\ paper2}]$$

(for the 2- and 3- parameter logistic models)

Hambleton and Swaminathan (1985) have constructed a four-step 'guide' for the equation of two tests using IRT techniques. The first two steps explain the need for the selection of an appropriate equating design and the use of an appropriate item response model. After these, the researcher is ready to establish a common metric for ability and item parameters. The most interesting thing is that if we calibrate all the items simultaneously, we have already equated the tests!

The result, according to Hambleton and Swaminathan (1985) and Lord (1980) is that we have already expressed all item parameters and all examinee abilities on the same scale. This is actually what we wanted - to create comparable results for every examinee - and this is what we obtained.

It should however be stressed once again in this report, (this is the professional opinion of the Consultant-which is in agreement with much of the literature), that the results of the joined calibration of the tests could be severely affected if the model's main assumptions are heavily violated. There can be many ways of violating these assumptions. A major violation of the unidimensionality assumption is caused by the different cognitive processes required in open-ended items or by using tests that cover very large areas of the curriculum.

The assumption of local independence can be violated-for example-by the presence of items that cluster around a common task (which could be more than one item depending on a single passage). These conditions were studied by many researchers like Ferrara *et al.* 1992 and Yen, 1992. On this issue, many researchers went on investigating the response dependency in tests with multiple-choice items, which are based on a single passage (Green and Langhorst, 1986; Hanna and Oaster, 1980).

This is a very important issue, and should not be taken lightly. Kolen and Brennan (2004) suggested '*For any equating design, the use of IRT methods requires making strong assumptions. Research should be conducted in the context of the testing program to make sure that the methods are robust to the violations of these assumptions which are likely to occur in practice*' (p.295). For example, Bolt (1999) evaluated the effects of multidimensionality on IRT true-score equating and reported that equity problems may arise when multidimensionality is present and the correlation between the dimensions is low, for example smaller than 0.7.

In the case of the IRT/Rasch models, the model-data fit is very important, especially in the context of high-stakes test equating tasks. The properties (for example, invariance) and the assumptions (for example unidimensionality and local independence) of the models are very important and should be investigated extensively. For example, Rupp and Zumbo (2004) published a paper to encourage researchers/practitioners to quantify (and suggest ways to report) whether IRT parameter invariance holds. They wrote:

> *There is no reason to believe that an excellent model fit in practice implies that parameter estimates are now valid for any set of items or any group of examinees from arbitrarily defined populations, which would be nothing but wishful thinking. On the contrary, it is always important to investigate for whom or for what items a given IRT model appears plausible to hold. Rather than believing that an IRT model invokes a proper test functioning across an infinite range of conditions because it is theoretically robust, one should rather ask whether such beliefs may mask an improper test functioning for certain conditions as the scoring methodology may be practically less robust for the conditions that one cares about.* Rupp and Zumbo (2004: p. 597)

In a very recent paper, Rupp and Zumbo (2006) argue that '*Parameter invariance is crucial if one wants to carefully assess the degree of inferential generalizability across examinee populations or measurement conditions for a given modeling context and thus constitutes a fundamental property of measurement for latent variable models*' (p. 64).

However, we should admit that, as Huynh & Ferrara (1994) and others showed, IRT models can be very robust to these violations, even when the model-data fit does not appear to be very strong.

Definition of validity and reliability in the context of test equating

A complete academic definition of 'test equating validity' will not be attempted in this report for the sake of brevity - it would be out of the scope of this report anyway. The Consultant is not aware of any research/academic work that deals with the issue of validity specifically in the context of test equating (maintaining standards), although one might come up with various definitions of validity, depending on his/her

background and experiences. Therefore, this report will draw extensively on general definitions of validity.

Reliability is a concept that is relatively easier to define compared to validity. In the context of test equating, Reliability could refer to the degree of reproducibility of the test equating results. In that sense, Reliability, in this report, is not considered as a dichotomous property (i.e. that either exists or does not exist); it is rather a matter of degree. How much reproducibility is enough in the context of English national curriculum test equating, might be a subjective matter; alas it may not be easy to reach a consensus. One rule of thumb might be to achieve a 'reproducibility of test equating results' that does not produce 'deviations' larger than the smallest block of measurement, which in this context is a single mark (i.e. level cut-off scores are set using whole marks and the raw scores are also reported in whole marks).

However, how could one define validity in this context? A test equating procedure should enjoy high face validity in the sense that it should 'look' right to the eyes of the stakeholders: i.e. the pupils, the parents, the teachers, the politicians, the psychometricians etc. Consequential validity is also of paramount importance, in the sense that setting the right levels, may affect the self-esteem of pupils, the evaluation of schools, the league tables etc. Other forms of validity might also be considered here, like concurrent validity which refers to the degree that the 'current' test equating results agree with the test equating results that might be obtained using other appropriate methodologies, methods etc. (note the subjective nature of the term 'appropriate', which in fact means 'valid methods' therefore resulting in a somewhat circular definition).

A special case in the literature review: The Massey report

In 2003, Massey and colleagues submitted a report which focused on the 'comparability of national tests over time'. This report, attracted the eyes of the Consultant immediately, but he decided to use it as a '*baseline*' on which to compare his own findings. The Consultant's assumption was that if he did not read the report, and worked totally independently, it would be possible to compare his own conclusions to the conclusions of the Massey report. It would then, in theory, be possible to identify areas of agreement and disagreement and draw further conclusions on the changes that may have happened (or have failed to happen) in the few last years.

Indeed, upon completion of the research, and when the Consultant had reached and recorded his conclusions, the Consultant read the Massey report very carefully. He identified a large number of issues that were of special interest. For example, the Massey report comments on the possibility to use different methods to carry out the test equating tasks and suggests that '*There should be good grounds for choosing one technique/result over another, or a defensible basis for compromise*' (p.235), and also argues that '*improving the quality of the statistical equatings brought to the table…[is]…of paramount importance*' (p.235).

In addition to the methods (i.e. models) used for the test equating task, the Massey report briefly discussed issues like the sample size of equating datasets, the pre-test effect, the strategy of equating back to a baseline year, dropped on the table the idea of external audits and even commented on marker 'drift'. For example, on the marker drift issue, although the methodology the Massey group followed on their experiment may have some drawbacks their results seem to be in agreement with the suggestion of Tate (2003) about marker 'drift' and the possible effects on test equating results.

In any case, the results of the Massey report are compared to the results of this report section by section (where the responses of the TDAs are presented and discussed). It is adequate, however, at this stage, to note that the Massey report mentions many issues which the Consultant is in full agreement with and these issues will be discussed in detail later in this report.

Statistical models

Different TDAs report that they use different models for test equating purposes. For example, National Foundation for Educational Research (NFER) staff use methods that are 'based mainly on classical test theory' (NFER, 2006), for example the equipercentile method is used to equate key stage 2 tests (science and English). In the case of the common pupils design, NFER developed its own software to carry out the equating and to estimate standard errors and confidence intervals for the equated scores. In the case of the equivalent samples, equipercentile equating is again used but no equating errors are reported. Smoothing techniques are also used where appropriate.

On the other hand, EdExcel staff (Hayes and Field 2006) believe that linear and equipercentile methods have specific problems and that the use of those methods in the context of national curriculum tests for test equating purposes may not be as appropriate as IRT. EdExcel staff use IRT methods for test equating purposes.

Cambridge Assessment staff on the other hand, currently derive the cut-off scores for different levels of attainment on different tests by simple linear equating or equipercentile equating using some form of anchoring.

It is appreciated that each of the above methods may be defensible, within a specific context, sample size, type of test, subject matter, practical intents and purposes etc. The question 'which is the best method of all?' may be the wrong question to ask, in the sense that there may not be a universally 'best' method that can be used in all contexts. However, an investigation into the issue will certainly help the interested parties choose the right tool for the right purpose.

For the above reasons, the consultant devoted the first section of the questionnaire to an exploration of the statistical models used by the TDAs for test equating purposes. The next sections present the relevant questions of the questionnaire, the responses of the TDAs and some comments by the Consultant.

Question 1.1

For which types of data (for example, English, mathematics, science, short response, essay, one-word response, multiple-choice etc.) do each model (or models) seem to be most appropriate/relevant to be used for test equating purposes (for example, dichotomous Rasch, 2-parameter dichotomous, 3-parameter dichotomous, Graded Response Model, Partial Credit Rasch model, linear models, equipercentile etc.)? Using which criteria?

TDAs' responses with comments from the Consultant

NFER staff suggested that in cases where only total score is available (for example, live test scores for pre-test pupils) the equipercentile is the most practical methodology for equating. They said that their experience of fitting IRT models to pre-test data has convinced them of the superiority of the 2-parameter graded response model for polytomous items over other approaches in terms of item fit and ease of interpretation. However, they still do not use any IRT models for test equating purposes, but they prefer the equipercentile equating. They avoided making any comparison between IRT and equipercentile equating in the case where item-level data is available, but again, they were not directly asked to make this comparison.

Consultant: The Consultant accepts the argument of NFER staff that when no item-level data is available, classical statistics seem to be the most practical solution. The same issue is raised by EdExcel staff as well; therefore more detailed comments will follow. Their appreciation for the 2-parameter graded response model for polytomous items contradicts the EdExcel practice where the 2-parameter logistic model is used, and any polytomous items are split in their dichotomous equivalents.

EdExcel staff suggested that if item-level data is not available (as is sometimes the case), classical statistics seem to be the most practical solution, though personal beliefs and preferences play a large part in the selection of statistical methods. EdExcel staff doubt that objective criteria could be found that would be accepted by all camps as evidence of model superiority. An example, say EdExcel staff, is that followers of Rasch models do not accept evidence that discriminations are unequal as a reason to introduce a discrimination parameter, arguing that the improved model fit is bought at the cost of violation of model assumptions.

Consultant: The Consultant shares the opinion of EdExcel staff to a large extent, in the sense that sometimes different 'camps' of analysts argue for decades about the merits or the shortcomings of their favourite statistical models. However, the statisticians usually do not disregard the statistical evidence put forward by competing camps; they rather interpret it in a different way or weight it differently. The Consultant, therefore, believes that it is in the best interest of the industry not to attempt enforcing some sort of guidelines or code of practice on the TDAs; it is better to reach a consensus on the issue of different equating models. A round table research meeting, where all the TDAs and QCA participate as equal partners with the facilitation of an external consultant, might achieve some progress. For example, it might be possible to set up a joint research agenda, which QCA might choose to partly or wholly fund, in order to investigate the statistical 'behaviour' of various competing equating models on existing, past and current datasets.

OCR staff explained that their preferred method of equating is the linear method, where no specialist software is needed. They also use an equipercentile method; again, no specialist software is used and the cut-scores resulting from the equipercentile method are not generally presented as their recommendation, but they are used as a second piece of evidence to use alongside the evidence from the linear equate. According to OCR staff, simplicity is one of the advantages of these statistical methods, although more complex methods might also be used if, as they say, they have evidence that they are more effective.

Consultant: The Consultant does not agree with OCR staff that simplicity can be a criterion to select an equating model. There is no need for the teachers, the parents or the politicians to know exactly how the equating model works; therefore there is no need to use a model that is simple enough for anyone to understand. What is required is to produce equating results that are robust enough and defensible, even if this means that we might have to use slightly more complicated techniques.

Issues not answered fully/clearly

The TDAs did not give clear answers as to which criteria should be used for the comparison between competing test equating models. Model-data fit, simplicity and availability of item-level data were the issues raised by the TDAs, but this might need

to be investigated further, especially in the case where a comparability study is to be conducted. Sample size requirements were also raised by OCR staff in question 1.2.

Suggestion

In a nutshell, NFER staff use the equipercentile model because it is practical, but they say that a 2-parameter graded response model has the best fit among the IRT models (but do not compare between equipercentile and IRT). On the other hand, EdExcel staff fully supports IRT models and actually use them (but not the 2-parameter graded response model endorsed by the NFER staff). OCR staff prefer simple methods so they use the linear and the equipercentile equating (though they prefer the linear over the equipercentile). OCR staff, however, say that they would not object to using IRT models if evidence of efficiency is given. So, the TDAs seem to disagree on almost all aspects of this issue.

Nobody, however, seems to object to using, or at least trying, IRT methods provided evidence can show that they are more or at least equally efficient. In the cases where the TDAs would prefer to use IRT, but they don't have item-level data (which seems to sometimes be the case, if one reads between the lines of the TDAs' responses), a way has to be found to provide item-level data to them.

In the cases where the TDAs do not agree which model is the most appropriate, a direct comparison between different models may be difficult to achieve. The Consultant knows this very well because he spent several months doing his Master's dissertation years ago on a technical comparison between linear, equipercentile and Rasch test equating, using key stage maths data from the 1998 testing cycle while he was a post-graduate student. At the time, the Consultant had also taken into account not only the flexibility and information provided by the models, but also the standard errors of equating (using bootstrap techniques or analytical solutions to equations where available).

The Consultant, however, is not aware of any recently published research between competing IRT models in the context of the national curriculum tests equating. There are large differences between the Partial Credit Rasch model, the Generalized Partial Credit model, the graded response model, the dichotomous 2-parameter model and the other IRT models that might be used for test equating purposes – some of these differences are well documented, for example see Sijtsma, and Hemker (2000). Some

papers even get into very large details, for example, see Baker *et al*. (2000). However, the industry might benefit if some research was funded to focus explicitly in the English context.

Moreover, as far as the preference for linear over equipercentile equating is concerned, it might be appropriate to conduct a small study using empirical datasets from the context of national curriculum test results in order to investigate the degree to which linear equating is indeed a satisfactory approximation for equipercentile equating. It is possible that the TDAs have already done research into this issue, with the findings in draft documents that could be published. If not Budescu (1987) might be a good starting point if one wants to investigate how well the linear equating approximates the equipercentile equating. Budescu assumes that sample sizes are adequate for both methods.

It is true that it would be very difficult for the TDAs to give a full-scale report in this document using specific arguments and examples to explain why they prefer one method over the other; it would inevitably be a very time consuming exercise. The Consultant accepts the argument by one of the TDAs that transition from one model to the other needs to be done with care, because it may lead to different equating results. However, this is exactly the point: if different models give noticeably different results, we need to know why we use the models we currently use, and we need to explain why we do not use the other models. On the other hand, if different models do not give noticeably different results (using English empirical datasets) then maybe we should stop worrying about this issue.

The above comment was also raised by Massey *et al.* (2003) who suggested that '*There should be good grounds for choosing one technique/result over another, or a defensible basis for compromise*' (p.235), and also argued that '*improving the quality of the statistical equatings brought to the table…[is]…of paramount importance*' (p.235). Since the issue of the different test equating techniques used by the TDAs is frequently raised by different people, our suggestion is that we need to (a) gather existing evidence (through literature review or existing unpublished research from TDAs) and (b) run an independent evaluation using linear, equipercentile and IRT models on empirical datasets (going back a few years maybe, because we need to investigate consistent conformity of the data to the models). There is surprisingly little published literature on this issue from England, thus, we may need to use (or analyse)

real data from our own context using the competing equating methodologies. As far as the criteria that could be used to compare different models are concerned, there is much relevant international literature: for example, one of the methods suggested by Harris and Crouse (1993) could be used to investigate the merits of one method over another.

Such a research could be undertaken by the TDAs, provided they are given a specific structure (a template?) in order to make results from different models as easy to compare as possible. An independent reviewer should also be involved in order to facilitate/coordinate the discussion/comparison. The task could take a peer-review format between the TDAs and academics might also be involved (the independent reviewer might also work as the coordinator/editor in the peer-review process). The TDAs should have direct involvement in the design of the aims of the study and they should have the opportunity to evaluate all the findings themselves, as a means of self- and peer-evaluation.

At the end of the exercise, the results of the study might be published on the internet, as well as published in more conventional forms, in order to share with the research community in England and across the world. This research might be used to show QCA's determination for scrutiny and transparency.

We need, however, to make sure that we do not force the TDAs to follow one of the methods; the TDAs should adapt their methods as a result of the research evidence. In the case where other TDAs move towards the linear or equipercentile equating methods, software may not be a problem. However, in the case where the TDAs move towards IRT, it is important to make sure that they use proper IRT software, and also take into account that there are various competing packages with advantages and disadvantages (see a relevant section later in this document).

Question 1.2

Using past experiences, relevant formal or informal research or publications, what are the minimum sample sizes for each of the tests, for each of the models? Has an interaction between sample size, statistical model and type of data been observed (for example, distinction between dichotomous items, essay, short response etc; or distinction between mathematics, English and science)? Are there any formal or informal reports, discussions, observations available on these issues?

TDAs' responses with comment from the Consultant

NFER staff explained that in many cases their sample sizes have been pre-determined by the sponsoring agency and they are not subject to negotiation. Their equipercentile equating software provides them with estimates of the equating standard error which is a function of the sample size as well as other factors, as they explained. This allows them to quantify the uncertainty in cut-score estimation based on finite samples.

EdExcel staff suggested that opinion varies as to the minimum sample size for various equating methods, and that anyone working in this field tends to have experience of a limited set of tests (compared to the range of tests in use). EdExcel staff suggested answering two questions before we can even begin to estimate required sample sizes: *what level of sampling variance are we likely to see from the test in question?* and *what level of accuracy is required from the equating process*? They argued that we should first ask whether our current sample sizes are sufficient by investigating whether the results have been good enough. They maintained that we do get acceptable results from the sample sizes we use. However, they said, if this were not the case, we could either increase the sample size (with all the resource implications that carries), or lower our expectations, or we could look for an alternative method.

OCR staff explained that for the English linear equate they ideally want to have a sample large enough to achieve a precision of better than 0.5 marks. They went on assuming a test standard deviation of around 15% of the marks available, and estimated that a sample of over 300 pupils per test version for English, and over 3,000 pupils per test version for science would be needed. They acknowledged that normally they have between 500 and 600 pupils per test version in their equating sample, for both English and science. As a result of this, OCR staff explain that '*it*

appears that we can achieve a precision of better than 0.5 marks for English, but not for science. With 500 pupils, the error in the mean for the science test will still be less than 1.5 (approximately 1.2), so our precision is still better than 1.5 marks, which means that when we recommend a cut-score of, say, 45, we feel it could just as likely be 44 or 46, but we are fairly sure it is not 43 or 47'.

OCR staff argued that they prefer the linear instead of the equipercentile equating because the linear method takes into account the score of every pupil (through the mean and standard deviation) to produce each cut-score; the equipercentile method only really takes into account the pupils on the marks in the region where each cut-score is, and is therefore more sensitive to small changes in the sample. The typical sample sizes involved are 600 pupils per test, corresponding to an average of twelve pupils per mark for KS3 English and only four pupils per mark for key stage 3 science. The sample sizes would need to be an order of magnitude larger for equipercentile to be the preferred method.

OCR staff went on to say that for key stage 3 science, although the preferred method of equating is that of linear equating, equipercentile equating is used to equate level 5 and level 6 cut-scores on the tier 3-6 to the equivalent cut-scores on the tier 5-7, using the items common to both tiers (although this linear extra equate is mostly for 'academic interest', and is rarely considered at level setting). The actual linear method used by OCR staff to equate science is known as the 'Tucker linear method'. This method uses pupils' scores from their live test to account for any differences in ability between the pre-test samples. This method is used because the way the science pre-test is administered does not guarantee equivalent groups (although the groups tend to be similar).

Consultant: One of the issues the Consultant would like to raise is that of the equating errors reported by OCR staff. An error of 1.5 marks for science at the mean of the test most probably means that the equating error is much larger off the mean of the test and especially at the tails (for lower and higher levels). So, the Consultant disagrees that such a precision is satisfactory, especially when the % of pupils affected may be too large at the borderlines of successive levels.

OCR staff also said that the linear equating needs a smaller sample size because it takes into account all the scores whereas the equipercentile takes into account only the

marks in the region where each cut-score is. Although this is generally correct, the linear equating could produce a substantially larger error because it implies a linear relationship that may not exist. In fact, the linear model may be considered as a special case of the equipercentile. Therefore, such a mis-specification, although perceivably produces smaller standard errors and is less sensitive to changes of the sample, could actually produce misleading results to the degree that we fit a linear relationship whereas a non-linear might be more appropriate. Budescu (1987) suggested a very good method to investigate the degree to which the linear equating is indeed a satisfactory approximation of the equipercentile. It is true that linear equating generally requires smaller sample sizes. However, Kolen and Brennan (2004) suggested that linear equating is most useful when the sample size is small, but also when the test scores on the two tests are not too dissimilar, '*…and a great degree of accuracy is needed only at scores that are not too far from the mean'* (p.292).

In addition to issues of linearity, in the case where the TDAs feel that their sample size may not be large enough for an equipercentile equating, then they might want to investigate increasing the sample size (with the all the resources implications that also need to be considered by QCA). A larger sample size might be used experimentally in one or two cases, in order to confirm empirically (if this evidence does not already exist in an unpublished report) that taking into consideration only lower distribution moments (i.e. linear equating) is not an oversimplification where higher moments might be needed (i.e. equipercentile equating).

Nevertheless, there is so much discussion about competing equating models that we would not expect this document to give definite and conclusive answers; a technical conference with all interested parties might be the best way forward. Also, as mentioned elsewhere, QCA might choose to offer secretarial, financial and other assistance to the TDAs in order to publish draft equating technical reports (especially comparative between competing models) and discuss practical equating issues.

The information given by OCR staff about the sample sizes and the possible equating errors may make one think whether the sample sizes of the other TDAs are satisfactory, although the other TDAs have not mentioned any sample size issues.

Massey *et al.* (2003) discussed the issue of final equating trials[1] and commented that '*…equating trials would require the assistance of sufficient schools to provide 1,000+ children in the required cohort for each test'* but they do not explain if they mean 1000+ pupils per paper. Under the light of their recommendation (they do not really explain where they base this estimated sample size of 1000), it seems that the current sample sizes are not always larger than this.

Issues not answered fully/clearly

The TDAs neglected to give information regarding the question: *Has an interaction between sample size, statistical model and type of data been observed (e.g. distinction between dichotomous items, essay, short response etc; or distinction between mathematics, English and science)?* The Consultant would expect the TDAs to discuss their experience with different equating models, sample sizes and subjects, but the TDAs may not have extensive experience with equating different subjects, using different equating models etc.

Suggestion

This question revealed some really important findings. The first is that the TDAs feel that sample sizes are (at least in some cases) pre-defined by the NAA. This, however, begs the question: *On what basis does the NAA pre-define the sample size needed*? The NAA may need to give a clear response to this question.

The second issue is that one of the TDAs acknowledged that some of their sample sizes may give an error (in the mean**)** of 1.5 marks. This however, makes one think that this error will probably be larger than 2 marks at the tails (it could even be larger). This begs for the question: '*What is, according to QCA, an acceptable equating error*?' provided the unit of measurement is 1 mark? We may also need to answer a further question: '*What % of students would be awarded a different level, had the*

[1] 'Final equating trials' could be built into test development schedules at little or no extra cost, about twelve months before each test's operational use, after the final version of each test had been fully and finally cleared by all government agencies. This would overcome two major weaknesses in current procedure: the imponderable effects of post equating changes required to the test and the unequal motivation of participants in current pre-tests towards the future test concerned and their own live version, to which it is equated. In the proposed trials, children would be taking either the baseline instrument or a future version of a test. They would not know which and would anyway have no reason to be more highly motivated on one than the other.

threshold been 1 or 2 marks up or down?', (this might indicate the sensitivity of the level setting decision taking into account the estimated equating error).

Finally, the responses of the TDAs to this question, inform us that the equating error may be different for different subjects. A final question we need to answer is: '*Is it acceptable to aim for different equating errors for different subjects*?' (Compare English and science discussion by OCR staff above). If the answer to the previous question is negative, we need to come up with a procedure so as to verify that the equating error for different subjects is comparable.

Question 2

Hayes and Field (2006) suggested that 'A further difficulty with equipercentile rating is that few pupils' score below the lowest threshold and therefore if any values can be obtained, they will be highly sample dependent. Effectively, this method cannot yield the lowest threshold on any of the tests' (page 4). Considering that NFER staff and Cambridge Assessment staff use the equipercentile method, it may be reasonable to ask if they have ever faced such a problem, and if yes, how was this problem overcome? Do the methods used by EdExcel staff avoid this problem?

TDAs' responses with comment from the Consultant

NFER staff commented that this is clearly a problem which is largely due to the paucity of data from pupils at these bottom ends of the scale, but it is not restricted to equipercentile equating methodology. NFER staff went on to say that other methods, including IRT, will run into problems when data is sparse, but may obscure this fact by extrapolation.

EdExcel staff argued that if one has very little data at the extremes of the mark range, then he/she is likely to be extrapolating the model beyond the limits of reliable data whichever method he/she choose to use. EdExcel staff went on to say that the pre-test data for mathematics has been getting progressively thinner around the level 3 threshold at both key stage 2 and key stage 3 but the IRT model is less dependent on having similar ability distributions.

OCR staff argued that the issue of having few pupils below the bottom threshold is one of the reasons they prefer to use linear methods rather than equipercentile methods.

Consultant: None of the TDAs confessed ever having the problem described in the question, i.e. to have a dataset that is too thin for a reliable estimation of a threshold here or there. This brings much relief, although the best practice might be to compare numbers rather than subjective statements. However, such an exercise would be beyond the specifications of this document.

Suggestion

The TDAs seem to generally agree that in the section of the scale where the sample is thin, the equating error gets larger (this is actually what the literature and our personal

experience says). One might wonder, however, why we don't boost the sample at the levels of the scale where we approximately expect the thresholds to lie. (This would not cause detrimental problems to our statistics, for example the mean, because we could weight the other parts of the sample/records). It is not clear how this might be practically done at this moment in time, this might be something to discuss with the TDAs (if it is at all practical or desirable).

This question revealed another area of disagreement between the TDAs. For example, a TDA commented that thin samples at parts of the scale encouraged them to use the linear method (implying that other methods do suffer more severely from this issue). Another TDA said that thin samples at parts of the scale could cause similar problems to all methods (presumably defending one of the methods here?), and a third TDA suggested that IRT may handle this issue in a more efficient way (probably defending IRT?). This is a problem that may need to be solved by research (it might be the only way to convince people to reach an agreement rather than forcing them to buy into one model). Such a research should be very well designed, should involve all interested parties and could cost a large amount of time and, possibly, money.

It is important at this stage to come forward and communicate to others how large the equating errors are at the areas where the samples are really thin (nobody presented any numbers about this). If there is a problem at some point, then it might be better to come forward and boost the sample sizes, which one of the TDAs has claimed are sometimes pre-defined.

Question 3

It has long been observed that pre-smoothing, as well as post-smoothing, techniques may be used/benefit a test equating task (e.g. Kolen, 1984, 1991; Livingston, 1993; Hanso et al., 1994). More recently, Kolen and Brennan (2004) provided a long list of pre- and post- smoothing techniques. On which basis did each of the English TDAs choose their preferred pre- and post-smoothing techniques (if any of the techniques are in use)? In the case where smoothing techniques are not used, are there any formal or informal studies to show that smoothing is not necessary in the specific context (especially taking into account the actual sample sizes used by the English TDAs)?

TDAs' responses

NFER staff said that a smoothing methodology has been adopted and implemented and appended Annex 1 and Annex 2 for further details. Both equated values and standard error estimates had been validated.

EdExcel staff do not use methods that involve the type of smoothing referred to.

OCR staff: do not use smoothing techniques.

Suggestion

It is not clear why NFER staff use smoothing techniques, whereas OCR staff (the other TDA that claims to use the equipercentile method in at least some cases) do not. Livingstone (1993) showed that a successful smoothing '*…reduced by at least half, the sample size required for a given degree of accuracy'* (p. 23), in a study where small-sample equating was explored. If there are certain advantages when smoothing is used, then the TDAs may need to communicate them to the other TDAs, in order to share good practice.

The effect of smoothing techniques (if used) on the results might be investigated more thoroughly, in the context of a possible future research (although there is much literature about smoothing techniques) using empirical data from the national curriculum tests.

Question 4

Are any of the Test Development Agencies studying the use of any other – maybe innovative – test equating methods? Are any evaluations in progress (either formal or informal)?

TDAs' responses with comment from the Consultant

NFER staff investigated the possibility of using kernel equating methods and also plan in future test development to make more use of IRT methods, especially to combine tests which are linked via common items or pupils – such as tiered tests, anchor tests etc. According to NFER staff, this procedure, using a 2-parameter graded response model, has proved eminently successful in the development of optional tests and will be transferred to statutory test development.

EdExcel staff said that their judgement methods are under constant review, as are their statistical methods.

OCR staff: No response.

Suggestion

In the future, one wonders whether it would be useful to include the graded response model or the kernel equating in a possible comparability study (research). It is also surprising that EdExcel uses the 2-parameter logistic model whereas NFER test the graded response model, so the disagreement between the TDAs is not only between classical/IRT methods, but there are different IRT competing models as well. At first glance, it seems that there is a divergent instead of convergent tendency as far as the equating models are concerned-thus making a comprehensive joint comparability study much needed.

The QCA needs to clarify at this point, whether they consider that research and development costs should be entirely covered by the TDAs (but presumably this would shift the cost back to NAA?). In any case, research and development are important issues, and a clearer strategy on this might be useful.

Would it be useful, if QCA encouraged the TDAs to develop a mechanism of sharing their research findings? Would it be feasible for the TDAs to establish joint research agendas and split the cost? Would a yearly journal published by either QCA or NAA be a proper means of encouraging joint or even separate research? How are the

responses to the above questions affected by the fact that there are commercial organisations involved in test developing? Test equating is surely one of the most developed psychometric/testing areas; however more research is always needed in order to tackle all the major issues that rise from time to time.

Data-model fit, assumptions and properties of models

This section deals with important issues that specifically relate to the IRT/Rasch world (although more 'traditional' methods like the linear and the equipercentile methods also have assumptions and properties which should be investigated and reported by their users).

Therefore, in our case, a number of questions arise in the sense that there are practically very few references (if any at all) in the reports of the TDAs on the properties and the assumptions of the IRT models. For example, the effect of multidimensionality on the equating results is not always addressed adequately in the reports of the TDAs. There are also very few references about the model-data fit. It is likely that if the model-data fit is not 'good enough' for our practical intents and purposes (this is, again, a qualitative judgment), then the equating results should be interpreted with extra care. It is recognised, however, that the model-data fit is always a matter of degree, it is not a have/not have property of the items (item fit) and the pupils (person-fit).

Question 1

Instead of (or in addition to) directly testing the assumptions of IRT models (for example, unidimensionality or local independence), one might choose to test the properties of item-free and sample-free measurement. For example, in order to test item-free measurement, researchers may split the test in two parts and run the IRT analysis for each; then they may compare the ability estimates of the candidates on each sub-test. This way, in the context of national curriculum test equating, one might test the effect of the anchor test on the ability estimates. The properties of the invariance (and therefore the robustness of the model to the not administered tests) could hold to the degree that item-free and person-free measurement holds. What formal methods/tests are used by the TDAs to investigate the assumptions and properties of the IRT models? How (if at all) can these methods/tests be presented in the test equating reports? How (if at all useful) can those methods/tests help the psychometricians evaluate the appropriateness of their IRT models for test equating purposes?

TDAs' responses with comment from the Consultant

NFER staff commented that this sounds like an interesting idea, though not one that they have investigated on their data. They said that extra resources would probably be needed to carry out such a study, as well as to collect additional data since half-samples may not be adequate for reliable estimation of parameters.

EdExcel staff argued that a calibration of items cannot be valid for any and every population of test takers and disputed the idea that a unidimensional age-independent scale can be constructed for mathematics, let alone any other subject. This is because some ideas remain hard however long you have known about them, while others become easier with familiarity. Inevitably, as they argued, this will have the effect of changing the difficulty order of a set of items across age cohorts. For this reason vertical equating across year groups is problematic, according to EdExcel staff, although it does not stop anyone trying. EdExcel staff went on to say that the issues of item-free and person-free measurement are important and that investigations into these areas can be very labour intensive and the results are not necessarily generalisable. EdExcel staff have, however, investigated drift in parameter estimates obtained from anchor tests, which are administered to different samples of pupils on successive years.

About the lack of detail in the level-setting report regarding assessment of model fit, EdExcel staff said that this is intentional and does not imply that such work is not undertaken. All TDAs, says EdExcel staff, attempt to provide reports that are fit for purpose and contain the level of detail as deemed appropriate by NAA.

OCR staff said that only the linear method assumes that the shape of the distribution of the anchor test scores, and the scores on the test being equated, are the same (although both methods assume that the ability of the samples are the same). They produce histograms of scores for each test, which are presented in the draft level setting report, and they look at all the summary statistics of the distributions to check that they are similar in shape… {the same response is presented for the next question as well}.

Consultant: The Consultant agrees that testing the assumptions of models (especially of the IRT models) could be a very time consuming task. However, the test equating is valid to the degree that the invariance property is valid, therefore not checking the

properties of the model means that we do not investigate the quality of the equating. Whether one wants to investigate the properties of his/her models is not an academic question, but mostly a practical and empirical one. Therefore, no matter how much it is going to cost, some basic tests of the assumptions and properties of the models are always necessary.

One of the important issues at this point, however, is that how one presents the results of such an analysis is frequently a matter of personal style and preference. This has forced academics and practitioners to issue guidelines on how to present the results of various analysis, for example the Consultant recently published a guide on how researchers should present the results of Rasch analysis (see Lamprianou, 2006), which is relevant to how one might want to present the results of an IRT analysis.

Moreover, the interpretation of the tests of the assumptions and properties of any model, and particularly the IRT models, is sometimes a subjective issue. However, as Kolen and Brennan (2004) have suggested '*For any equating design, the use of IRT methods requires making strong assumptions. Research should be conducted in the context of the testing program to make sure that the methods are robust to the violations of these assumptions which are likely to occur in practice'* (p.295). Since different researchers could reach slightly different interpretations using the same results, it is always a good idea to present the results and let the others draw their own inferences. Having said this, a technical report could have a different audience than a report to facilitate setting the standards aimed for teachers or policy makers.

Suggestion

The first issue that rises from the responses of the TDAs to this question is that not all TDAs seemed to be so enthusiastic about the concept of investigating the assumptions of the models, (neither have all of the TDAs produced research in this area). However, it seems (from their responses) that they consider this issue to be important, and we would definitely like to see more research on this issue.

It is generally agreed that this type of research is costly (both in time and money wise). It is not clear whether, or how, we could reach an agreement on routinely doing some basic checks of the assumptions of the models (it is not clear if this is even desirable for all of the TDAs), although investigation of the assumptions of models is always a

useful task. However, would this task be too 'academic' for the practical purposes of the TDAs?

Finally, the TDAs apparently do not publish much of this information. There was a very important comment by one of the TDAs: '*The lack of detail in the level setting report regarding assessment of model fit is intentional…In the past we have not been thanked for providing more information than was wanted or asked for…*' Although this is something that might be more appropriate to discuss in a section later in this report, it seems that QCA did not consistently encourage the TDAs to give more information, or to do some more analysis of their data. At the least, the TDAs might want to charge QCA for producing these technical reports.

Question 2

There are a large number of statistical tests by which the model-data fit can be evaluated for IRT models. Which are the models used by the TDAs to evaluate data-model fit? What formal procedures do the TDAs use in order to judge whether the data-model fit is 'good enough' for the practical intents and purposes of the test equating? Are there common criteria across the TDAs by which to judge the quality of the model-data fit (in other words, would all TDAs take the same or similar decisions upon evaluation of the same data-model fit)? Has a situation ever happened, when the model-data fit was not satisfactory? What is the contingency plan in such a case (for example, will a different statistical model be used)? How is model-data fit presented (or should be presented) in the test equating reports?

TDAs' responses with comment from the Consultant

NFER staff: IRT is not used for test equating – although they could have commented on the methods they use, as OCR staff have done.

EdExcel staff mentioned that there are no agreed processes for judging model fit, across the TDAs (although they do tend to use similar ideas). As pointed out by EdExcel staff, different programs report misfits in different ways making rationalisation of methods of assessment difficult. The 2-parameter model, according to EdExcel Staff, will generally fit the data better than the 1-parameter model so a straight comparison would be meaningless and criteria for judging acceptability might need to be different for the two models.

OCR staff stated that only the linear method assumes that the shape of the distribution of the anchor test scores, and the scores on the test, being equated are the same. They produce histograms of scores for each test, which are presented in the draft level setting report, and they look at all the summary statistics of the distributions to check that they are similar in shape…{the same response is presented by the Consultant for the previous question and for question 6 as well because it is relevant}.

Issues not answered fully/clearly

The TDAs did not indicate whether they ever faced a problem with model-data fit. Although this is refreshing, we still do not know whether there are any contingency plans in place in case the model-data fit is not satisfactory.

Suggestion

It is apparent from the responses of the TDAs that they do not use similar methods to evaluate data-model fit. It is now clear that agreed methods between the TDAs do not exist. From this perspective, we need to answer the question: *Is it desirable for the TDAs to agree on some basic/common principles on these issues*? First, and then to try to see how we can come up with mutually agreed and acceptable procedures/methods.

Most importantly, however, the TDAs did not deal with the following questions: *Has a situation ever happened when the model-data fit was not satisfactory? What is the contingency plan in such a case? (for example, will a different statistical model be used?).* In fact, we all know that the national curriculum tests are of good quality, and we would normally expect the data to fit the usual psychometric models, for example IRT to a satisfactory degree. However, and this is the most important, the TDAs did not answer the question about their contingency plans: *What is the contingency plan in such a case? (for example, will a different statistical model be used?).*

This issue may not be important if the TDAs have never faced any serious problems with their statistical models, as seems to be the case.

Question 3

There are no indications at all in the test equating reports that the person-fit has been considered. Person-fit is an issue that is rarely investigated but large person misfit for specific sub-groups may be a strong indication that these sub-groups have been mis-measured. Lamprianou and Boyle (2004) showed that certain sub-groups of pupils may be consistently mis-measured in the context of the national curriculum tests in England. To the degree that a sub-sample of pupils is consistently mis-measured, and depending on the representation of this sub-sample in the equating sample, how might this endanger our test-equating results?

TDAs' responses with comment from the Consultant

NFER staff said that person-fit is not an issue for the equipercentile equating, which is essentially non-parametric and based on individuals' total test scores. They acknowledged that it might be an issue if equating results were significantly different for different sub-groups but confessed that this has not been investigated, as the purpose of the equating is to define a single set of cut-scores for all pupils, based on a suitable representative sample. More research could be carried out in this area, NFER Staff said, if QCA felt this to be a fruitful use of resources.

EdExcel staff mentioned that '*different programs report misfit in different ways making rationalisation of methods of assessment difficult. The 2-parameter model will generally fit the data better than the 1-parameter model so a straight comparison would be meaningless and criteria for judging acceptability might need to be different for the two models'*{this is an extract from a response to a different question}.

OCR staff '*admitted that they only report on the mean and standard deviation, as these are used in the calculation of the cut-scores'* {this is an extract from a response to a different question}.

Suggestion

The TDAs may not think that this issue is very important. In any case, it seems to be the case that no research has been conducted recently along those lines, so we would not know whether this is still a problem; this was probably a problem, at least academically, a few years ago according to Lamprianou and Boyle (2004).

However, Kolen and Brennan (2004) warn against this type of threats and suggest carrying out '*robustness studies'* like '*…equating can be conducted for various sub-groups of examinees... to the extent that the equating is robust, the equating should be similar in the various subgroups*' (p.297). It may be true that the TDAs do not worry too much about this issue; however, it seems that other experts advise that we should investigate those issues more closely. Such research could be time consuming, and QCA might want to contribute towards the cost.

Question 4

What is the effective size of the mis-specifications of the model? Are there any simulations or other studies about this?

TDAs' responses with comment from the Consultant

NFER staff explained that this is not an issue for their methods because they are non-parametric, though they said that studies of the dimensionality of national test data might well be interesting and give additional insights into this area.

EdExcel staff reasoned that they are not averse to running simulations and have undertaken several to reassure themselves as to the effectiveness of their methods. They view this as research/professional development – as such it would not be appropriate to report outcomes to NAA/QCA. However, they would be pleased to be commissioned to undertake specific pieces of research.

OCR staff's preferred method, linear, assumes that the skew and the kurtosis of the distributions are the same. Most of the time this has been the case, to a reasonable approximation (i.e. the difference between the distributions is less than two standard errors for both the skew and the kurtosis). When there is a large difference between the distribution, the equipercentile method is also considered (but not preferred – the sample size used is not large enough for the equipercentile method to be reliable).

Consultant: Simulations have served the psychometric community for many years. Especially in the context of test equating, Harris and Crouse (1993) suggested using the '*simulated equating'* as a criterion to investigate the merits of one equating model over another. Harris and Crouse also suggested the '*large sample criteria'* which might be considered a form of simulation: data from a very large sample of examinees are used to represent the population and then drawing from this large dataset for test equating purposes may '*simulate*' an empirical test-equating task. However, the Consultant appreciates the concerns of some TDAs that simulations may, in some cases, work as a self-fulfilling prophecy. Careful simulations, however, in conjunction with analyses using empirical data, might give valuable help in solving issues discussed in this document.

Suggestion

The TDAs may not consider this issue to be very important: maybe this question is too ‘academic’ and does not fit with the practical aims or concerns of a TDA. However, the choice of a linear over the equipercentile equating, is a possible misspecification issue which is at least interesting, if not important.

Question 5

Is there a guideline from QCA to include in the statistical reports indicators of the reliability of the statistical information, for example standard errors, item fit, especially the fit of the anchor test/item etc.

TDAs' responses with comment from the Consultant

NFER staff said that the more information about standard errors and item fit etc. that is included the better. They normally include confidence intervals from their equipercentile equating, and would expect that in reporting IRT parameters. Measure of fit would also be included, as well as total test information functions and similar information. They have not commented on whether they feel that QCA's guidelines require this information to be reported.

EdExcel staff gave no response here, but elsewhere they specified that they had not been encouraged to include additional information in the reports: '*In the past we have not been thanked for providing more information than was wanted or asked for. We will always raise any concerns we have over model fit with NAA and advise on what steps we think are necessary to deal with the situation' {also cited in a previous section}.*

OCR staff: gave no response here.

Issues not answered fully/clearly

None of the TDAs actually answered the question, i.e. if QCA's guidelines require this information to be reported.

Suggestion

The TDAs may not consider this issue to be very important since they did not provide any specific comments, (or probably were not sure what exactly to respond to in this question). If there are truly no guidelines from QCA, is this because QCA does not really think that this is important/necessary? This needs to be clarified by QCA.

It would also be interesting to see whether the TDAs feel that this is an area that should be regulated. During informal discussions with TDAs' staff, we got the impression that too much regulation in this area might be counter-productive.

Question 6

Also, are there any explicit investigations of the assumptions of the equipercentile or linear methods (by those that use them)? Not much statistical information on this issue is presented in the reports.

TDAs' responses with comment from the Consultant

NFER staff mentioned that equipercentile equating is essentially non-parametric and makes few assumptions, apart from the main one that both tests are taken by the same pupils (or equivalent groups) under the same conditions, including motivation etc. The last part is violated when they equate live and pre-test scores, but this is explicitly highlighted in their discussions of 'pre-test effects'. When equating via equivalent groups (pre-test one year to pre-test the next) they weight by achieved test levels. This implicitly assumes that levels one year convey the same standard as levels the next year. Occasionally they compute linear equating lines at the same time as equipercentile, but use the latter for level-setting as they provide more detail.

EdExcel staff: Not applicable – the question explicitly refers to the classical methods.

OCR staff said that only the linear method assumes that the shape of the distribution of the anchor test scores and the scores on the test being equated are the same (although both methods assume that the ability of the samples are the same). They produce histograms of scores for each test, as they said, which are presented in the draft level setting report, and they look at all the summary statistics of the distributions to check that they are similar in shape…{the same response is also presented for previous questions by the Consultant, because it is relevant}

Consultant: The issue of which equating model is used is not academic, but a purely practical one. For example, Kolen (1981) clearly stated that '*…conventional equipercentile or linear equating can be strictly used only with parallel tests… In theory, Item Response Theory models…can be used to equate both parallel and non-parallel tests…*' (p.1-2). Therefore, the issue of which models we use, and how we investigate their assumptions (for example the assumption of parallel tests in the case of classical equating methods-an assumption rarely mentioned by the TDAs in this questionnaire) as well as why the different TDAs use different models to equate their tests, should be investigated further.

Suggestion

The TDAs seem to have done some investigations of the assumptions of their models. Maybe it would be more appropriate to include this information in their reports, since they already spend their time carrying out the analysis, but they should be encouraged and probably be given some credit for doing so.

The quality of the datasets/samples

The reliability of the marking, the sample sizes, the sampling procedures and other similar issues are discussed here. The sample size is not the only important aspect of data that could affect the test equating results. Issues that have to do with the quality of the sample e.g. representativeness, quality of the marking (especially of the anchor scripts) are also important.

Question 1

What procedures are in place in order to verify that the quality of marking, especially of the data used in the equating procedure, is of high standard?

TDAs' responses with comment from the Consultant

NFER staff explained that their markers are experienced and are trained to mark by members of the project team using mark schemes which have been refined by the team during pre-marking trials. They added that the marking is checked on a random sampling basis and that consistent errors are identified and dealt with by retraining the markers and/or re-marking as appropriate. Indeed, NFER staff attached a very informative and elaborate piece of research by Tom Benton titled *Exploring the importance of graders in determining pupils* examination results using cross-classified multilevel modelling" which is a very good example of what TDAs could do in order to evaluate the quality of their operational marking. However, the Consultant would argue that such a complicated piece of research would be impossible to routinely carry out in a formative manner in order to monitor the marking procedure in 'real time' as we would say. More on this issue, however, are mentioned below.

EdExcel staff explain that they always hire highly experienced markers, actually aiming to use the most senior markers available (generally at least of Team Leader standard for the test in question). In addition to this, all markers have specific training for the papers they are to mark. Although monitoring the marking is not explicitly mentioned in the response, it is likely that this is happening in an organised way; (the question set by the Consultant could have been clearer, asking explicitly more details about the how the marking is monitored by the TDAs.)

OCR staff also mention recruiting senior, or very experienced, markers who attend training meetings at OCR, and mark training/coordination scripts. OCR staff went on to say that the marker training meetings for the second pre-test (Science) and the marking pre-test (English) are run by the Marking Programme Leader and Deputy Marking Programme Leaders. Monitoring the operational marking was not explicitly mentioned in this response. The issue still needs to be discussed and clarified, since the monitoring of the marking process (as well as the quality of the marking) is very important.

Further discussion of TDAs' responses

All three TDAs explained that they use various methods to make sure that the marking of the scripts used for the test equating is of high quality. All TDAs use very experienced markers who are trained and their performance is presumably monitored in an organised way. However, the Consultant, after spending years investigating marker effects in various countries (for example, Malta, England, Cyprus, Greece) does not believe that what is currently done in England guarantees the best possible quality of marking, not even after taking budget issues into account.

Understandably, monitoring the quality of marking for hundreds of thousands of scripts (when the scripts of the whole cohort are considered) is a very difficult thing to achieve. However, monitoring the quality of marking for the few hundred scripts used to make up a test equating dataset is easier to achieve, cheap and also desirable. It is easy-research shows that it happens-even for the most experienced markers to be out-of-tune with the rest of the markers and this is especially important in cases where a single marker could mark 10% or 20% of the scripts that make up the equating dataset. In a recent unpublished study carried out by the Consultant in Cyprus, a senior language marker was one of the most unreliable and the most lenient markers from a group of almost 60 markers. Although he participated in the moderation and although he marked ten 'co-ordination' scripts before the operational marking, by the time his marking was statistically evaluated (and judged to be problematic), he had already marked almost 200 scripts. What caused this unusual marking behaviour is not of importance at the moment (although this was investigated fully), what is important is that this could also happen in the context of the national curriculum tests.

One might argue that in the case where a small number of markers mark the scripts that make up the test equating samples, then the effect of a single marker being 3-4 marks more lenient or harsher than he/she should be, could lead to practically significant test equating effects. Massey *et al*. (2003) presented results suggesting that four experienced markers who had the same training procedure, used the same mark scheme and the same co-ordination scripts had up to three marks difference in the average scores they awarded when scoring the same 24 key stage 3 English Paper 1 scripts (approximately ¼ of a level).

Suggestion

The Consultant suggests that the scripts to be used for test equating purposes should be double-marked blindly (so that the first marker would not be aware of the scores of the other markers)-if this is practical. The Consultant understands that this may have specific practical problems-probably problems that cannot be solved.

If possible to do, double-marking might be useful in the case of English and science, although the mathematics tests might not need double-marking. Double-marking only the scripts to be used for the test equating will not be so costly and will not delay the procedure too much because of the small sample size. It is also suggested that the quality of marking be checked statistically by using the appropriate statistical models, for example multi-facets Rasch models.

This issue is also related to the next question.

Question 2

In the case of anchor tests, is it possible that the parameters of the anchor data might be different in the next years if the responses of the examinees were re-marked again (for example,. in case that the markers' standards/marking behaviour changed across time). Such a thing may affect the test equating results significantly. Tate (2003) suggests a solution:

'The proposed method is a modification of the traditional common-item nonequivalent groups design. Rather than collecting equated item parameters from previous tests, the method would require the collection of actual examinee responses from previous years and the rerating of those responses in a linking study conducted for the current equating. The resulting design controls for any year-to-year changes in the standards of the rating team, isolating the effect of any change in the examinee ability distribution in the computation of the equating transformation'. Has this issue ever been considered, and if yes, has any relevant research been done?

TDAs' responses with comment from the Consultant

NFER staff suggested that a remarking exercise on old scripts is not carried out, and that such an activity would be difficult on practical grounds, i.e. as the marks awarded are written on to the scripts they cannot be efficiently and effectively removed on the scale required. In other words, it will be difficult to produce 'clean scripts'. Indeed, such an issue has been raised occasionally by other researchers as well and the Consultant agrees that it may need to be addressed some way. Other countries, for example, Cyprus and Greece have solved this problem many years ago by providing a specially designed space on the front cover of the examination scripts where the markers write the scores awarded per question/item. This has the advantage that the scripts remain always clean, double-marking is easy and in the case of appeals the re-marker will not see the scores of the initial marking. In addition to this, further research by re-marking or photo-copying the scripts is made easy. Finally, the new technical advances, for example on-screen marking could solve this problem very soon in any case.

EdExcel staff on the other hand, explained that at key stage 3 they use papers from the previous cycle as anchors so the time frame for considering variation of item

parameters is limited. They went on to suggest that key stage 2 presents a better opportunity to review the stability of item parameters over time, since the anchor test remains the same for a number of years. However, the likely effect sizes in mathematics, they argued, are probably very small since the mark schemes are so detailed and prescriptive. EdExcel staff concluded that '*…we would not expect any systematic difference in severity over time and do not advocate the design proposed by Tate (2003')*'. In the case of such an exercise, EdExcel staff suggest that a very careful design would be needed to avoid the introduction of more variance than it was intended to control.

Consultant: The Consultant agrees that mathematics is a subject which favours neither much discrepancy between markers nor differential marking harshness across time. The reliability of marking research usually focuses on more 'subjective' subjects such as languages, social studies and history, for example see Lamprianou (2006) where the stability of marker characteristics across subjects is discussed. Therefore, the issue raised by Tate (2003) may be more interesting to those developing tests for English and science.

EdExcel staff (cont.) In addition to the above, EdExcel staff discussed the issue of item drift in cases where tests are taken by successive cohorts and argued that there is a need to investigate possible causes where '*…if an intervention such as the National Numeracy Strategy produced an overall improvement in performance…[or by]…over-exposure of the anchor items or changes in emphasis between the teaching of content on the anchor test and teaching of the content of the main test'*.

Consultant: However, neither Tate (2003) nor the Consultant were referring to the item drift in the usual meaning of the term; this question rather focused on the differential harshness of the markers involved in successive operational marking cycles. Such a differential harshness of whole groups of markers could arise because of different expectations of the markers/teachers – a very good recent example of an empirical study in the context of the National Curriculum tests is given by Massey *et al.* (2003) is discussed in more detail below. The issue, however, of item drift is indeed a very interesting one and may need to be discussed further.

OCR staff acknowledge that the nature of English and science, and the format of their papers, means that marker reliability is more of an issue for English than for science

at key stage 3. They explicitly clarify that additional procedures are in place to ensure the quality of the marking for KS3 English. For example, all markers are asked to mark ten standardisation scripts half-way through their marking. In addition to this, each marker marks the same ten scripts, which have previously been marked by the OCR staff team, presumably enhancing consistency. Specifically for English, they explained that the key stage 3 English anchor test is marked at a different time of year to the test being equated, so they are able to use the same markers for both sets of marking, and they aim to have exactly the same markers marking both tests.

Consultant: The Consultant thinks that the above arguments are not convincing and do not address fully the issue raised by this question, which specifically targets the differential harshness of the markers across time when anchor tests are involved.

Further discussion of TDAs' responses

The TDAs do not seem to think that the issue raised by this question is either important or worrying. However, besides Tate's (2003) academic concerns, there is some recent research evidence, in the context of the national curriculum test results, that should also be considered. For example, Massey *et al.* (2003) carried out an experiment where the 1996 and 2001 version of key stage 3 English Paper 1 were administered to hundreds of pupils. This was phase 1 of the experiment. Four markers marked the scripts and it was found that the pupils received, on average, 0.27 levels higher in 2001 compared to 1996. However, just after the four markers finished marking phase 1, the second phase of the experiment took place, where the same four markers were given 24 key stage 3 English Paper 1 scripts from the 1996 live marking exercise (they were 'clean' copies of the scripts). In this second phase, the raw scores awarded by the four markers were on average 4 marks lower than the actual scores the 24 scripts received back in 1996. According to the authors, 4 marks may be estimated to 0.27 levels, therefore, exactly cancelling out the gain measured in the first phase of the experiment. They concluded that '*… today's markers would be likely to award lower marks than those marking operationally in 1996…Our best estimate of the extent of any under-marking involved is…about four marks…0.27 of a level…*' (p. 93-94). The research concluded that markers' standards may not be stable across time even when marking the same scripts after having the same briefing, the same mark schemes and the same exemplar 'co-ordination' scripts. Under the light of this evidence, even if the methodology of Massey *et al*. may be challenged on many

grounds, it seems that the suggestion of Tate (2003) about marker drift (in either direction) may not be out of order, and that it might, actually, be worth some more attention.

Suggestion

It is suggested that this issue is raised once again, and possibly discussed with the TDAs, taking into account the new information offered to them (not existing in the original questionnaire) by the Massey *et al*. (2003) research. If the TDAs still believe that this is not an important issue, then the Consultant would suggest to QCA to follow the advice of the TDAs, since they are in a better position to know whether their markers might be affected by a 'marker drift' or not. In the case where the TDAs would agree to look into this issue more closely, it is possible that funding might be provided by QCA in order to design and carry out a relevant research. In any case, if a relevant research is conducted, this should involve the TDAs in the formulation of the aims, in the planning, as well as in the actual conduct of the research. Although this might not be an issue for mathematics, the problem (if a problem exists) could be more apparent for English (and less probably for science). It is also advised that some external evaluation of the study be sought during the planning phase and for the interpretation of the results.

Question 3

Test equating needs to comply to the relevant 'Code of Practice'. How is compliance with the 'Code of Practice' monitored by the TDAs themselves?

TDAs' responses with comment from the Consultant

NFER staff explained that details of the equating methods used are supplied in draft level setting reports and different methods are compared as appropriate. According to NFER staff the reports comply with section 2.4 of appendix 2 of the Code of Practice and all reports are checked by the Statistics Research and Analysis Group and by the assistant director responsible for assessment and measurement, or his nominee, in order to ensure compliance with the Code of Practice.

EdExcel staff said that all the work they do is tightly monitored by them (the analysts presumably), their Programme Office and by NAA.

OCR staff: No response.

Suggestion

The responses from the TDAs could raise the issue of the internal monitoring of compliance, in the case where no clear responses were given to the question.

Question 4

Related to the above, in the Code of Practice there is a specific requirement: "This sample should include: e) an equal number of boys and girls; f) pupils for whom English is an additional language; g) pupils who have special educational needs. Should there be efforts to include proportionally in the sample the % of pupils with SEN and EAL? What will be the effect if SEN or EAL pupils are substantially under- or over-represented in the sample?

TDAs' responses with comments from the Consultant

NFER staff said that although when sampling they include a separate non-representative sub-sample (as required by NAA) to look at differential item functioning for pupils for whom English is an additional language, this sub-sample is normally excluded when performing the main analyses including any equating. Their aim is to ensure that the equating sample is as far as possible representative of all groups.

EdExcel staff said that they do not generally acquire data on gender, SEN or EAL until after schools and classes have been selected for inclusion in the sample. Efforts to include the appropriate proportions are not, therefore, guaranteed to be successful. The options, say EdExcel staff, are to re-weight data or select cases to provide the 'correct' balance in a smaller sample. However, they know from comparisons with figures held by the Department for Education and Skills (DfES) that the data on SEN and EAL are under reported, but not by how much. According to EdExcel staff, pre-test sample sizes are far too small for effective analysis of DIF even if they had high quality data on EAL or SEN.

OCR staff said that for both key stage 3 English and key stage 3 Science, when schools are recruited to take part in the second pre-test, they aim to recruit about 5 schools with a high proportion of EAL pupils. They do not specifically recruit SEN pupils, but they expect a representative proportion of them in their sample. But they cannot provide modified tests (as these are not produced until the end of the test development cycle) and they cannot allow pupils to use computers, as this compromises the security of the test. They believe that with their model of equating it should not matter greatly if SEN and EAL pupils are under- or over-represented, providing they are represented to a similar extent in the two samples. However, if

EAL and SEN pupils are substantially over-represented, this may have an effect on the overall ability of the sample resulting in, for example too few pupils at Level 7.

Consultant: The TDAs do not seem to think that this is a major source of concern. However, the following issues arose:

(a) The equating sample sizes are too small for a satisfactory DIF analysis.

(b) Even if the sample sizes were larger, there seems to be some practical problems in getting samples with proportional representation of the sub-samples, for example SEN, EAL etc.

Suggestion

At the moment, the Consultant is not sure whether QCA should invest money or time pursuing this line of inquiry, since the TDAs do not feel that this is an issue and there is no strong evidence that they are wrong.

Question 5

What about monitoring the relationship between item score and total score over time to monitor for item drift? Have any of the TDAs done any research on this issue? Would it be practical to use this in practice?

TDAs' responses with comment from the Consultant

NFER staff said that since new items are written for each year's test, this is only relevant for the anchor tests, but item analysis of the anchor tests is carried out each year, and significant changes in item parameters would be flagged. In practice, whole anchor tests tend to be replaced at regular intervals because of changes to the curriculum and/or style of testing.

EdExcel staff have looked at the relationship between item score and test score over time by examining time series data on the *a* (discrimination) parameters for items on the key stage 2 anchor test. The practical use of this is to highlight items that are losing their discriminatory power over time. This is usually due to one of two reasons:

1 Familiarity with the question type has risen to the extent that virtually all children answer it correctly.

2 The topic is no longer part of the programme of study.

In both cases, the item should be reviewed for replacement, say EdExcel staff.

OCR staff suggest that the anchor test items should be monitored for item drift, especially when they have been used for several years. The key stage 3 Science anchor test is relatively new (first used in the 2006 development cycle). The key stage 3 English anchor test is a little older (first used in the 2004 development cycle).

Suggestion

How and whether a common policy from all TDAs should be pursued in the context of item drift is not certain. The TDAs will certainly have their own views and approaches and it might be appropriate for QCA to consult them on this issue. Sharing good practice between the TDAs is a very good idea and will help the industry. However the fact that commercial companies may be involved could complicate things a bit more. It is up to QCA to decide whether item drift research/guidelines are something that should be pursued in more depth.

Test equating design–test equating error

The test equating design is the topic of this section. The design of test equating is of paramount importance both for the validity and the reliability of the results. Taking into account the fact that any TDA could follow one of many test equating designs, this section is slightly difficult to investigate.

The issue of test equating design has been discussed in the Literature Review section and examples and relevant tables have been presented.

Question 1

One of the most important issues in this topic is the so called 'pre-test effect'. This is documented in the international literature, although not widely addressed in a large number of publications. There is, however, much unpublished discussion about this in England. Are there any studies by the TDAs that try to quantify the pre-test effect?

TDAs' responses with comment from the Consultant

NFER staff explained that pre-test effects tend to be thoroughly discussed in the level-setting reports for the various subjects each year, with incremental information being built up over time. However, according to NFER staff, the scale and nature of the pre-test effect seems to vary between subjects and in English the evidence is that it is relatively stable over time, of the magnitude of 2-3 marks. For science, NFER staff said that the pre-test effects tend to be larger (up to about 10 marks) and more variable over time and across the score range.

More exact quantification of the pre-test effect could be achieved in various ways, according to NFER staff, mainly by linking items taken under live test and pre-test conditions using IRT models. One relatively cheap way, says NFER staff, of investigating this effect would be to capture item-level data from the live tests and combine it with pre-test data on the same items. Research projects along these lines could give useful information.

EdExcel staff gave a positive response, saying that they have tried to quantify the pre-test effect but they have not been successful. They recognised, though, that performance tends to be lower in pre-tests than in live tests and that there are several factors involved:

1 Motivation-live tests 'count'; pre-tests don't.

2 Preparation-the live test is the focus of preparation; pre-tests are inevitably 'off peak' even if only by a week or so.

3 Curriculum effects.

4 Environment-school hall or classroom?

5 Timing-the gap between components; morning or afternoon?

6 Time allowed-pre-tests might be squeezed into a particular session.

7 Test fatigue-if the pre-test follows the live test.

.

Edexcel staff feel that unless they are in a position to identify which are the important factors, and build them into their model, analysis of the data is unlikely to be sufficient to explain the phenomenon.

OCR staff say that there have been studies that try to quantify the pre-test effect, but it is very difficult to quantify accurately, because there are so many variables (year, time of year, school, venue, i.e. hall/classroom). For this reason it seems to OCR Staff most sensible to use an equating design where the pre-test effect is the same for the test being equated as for the anchor test (so the pre-test effect should be cancelled out). This, according to OCR staff, is what they aim to achieve with their model, where the anchor tests are taken in the same pre-test as the tests being equated, and should therefore be subject to the same pre-test effect.

Consultant: All three TDAs consider the pre-test effect to be very important. They all claim to be doing some research on it, with varying degrees of success, presumably based on various criteria. However, it seems that they all agree that more research is needed on this issue, though there is no consensus on what should be done and what strategy should be followed on the issue.

The pre-test effect is not new, it has been widely discussed in the past and recent reports suggested taking it seriously into consideration when carrying out the test equating tasks. This is also presented in the section below.

Suggestion

The so called pre-test effect is a serious problem which concerns all TDAs. Some investigation has already been done by the TDAs.

This problem may be too complex for a single TDA to research with its own resources. Some coordination, and probably joint research, between TDAs might be more efficient, not only because economics of scale would save money, but joint research could bring the researchers of the TDAs closer to adopting similar strategies and good practices.

It might be a good idea to ask TDAs to provide QCA with formal research proposals aiming to measure the pre-test effect. The TDAs would probably need support from QCA on this aspect, in order to try to give some definite answer on this issue (if this is possible). Joint research proposals could be encouraged by additional funding.

However, besides measuring the pre-test effect (which could be variable from year-to-year, across subjects, gender etc.) more research is also needed to devise new test equating methods (or to evaluate existing ones) in order to identify methods least affected by the pre-test effect. For example, Massey *et al*.(2003) suggested a method to overcome the pre-test effect which causes '… *unequal motivation of participants in current pre-tests towards the future test concerned and their own live version, to which it is equated*…' (p.233). They actually suggested implementing a Final Equating Trial which:

> *' could be built into test development schedules at little or no extra cost, about twelve months before each test's operational use, after the final version of each test had been fully and finally cleared by all government agencies. This would overcome two major weaknesses in current procedure: the imponderable effects of post equating changes required to the test and the unequal motivation of participants in current pre-tests towards the future test concerned and their own live version, to which it is equated. In the proposed trials, children would be taking either the baseline instrument or a future version of a test. They would not know which and would anyway have no reason to be more highly motivated on one than the other.* Massey *et al.* (2003, p.233)

The Consultant would like to see this issue being discussed with the TDAs, if it has not already been discussed.

Question 2

How different are the results of the pre test-pre test and live test-live test equating? Has this been quantified by any research by any of the TDAs?

TDAs' responses with comment from the Consultant

NFER staff said that the pre-test to pre-test equating is a 'valuable element' in their armoury of equating methods (using equivalent groups, as discussed above), but '*it is not clear how live test to live test equating would be operationalised'*. NFER staff suggested that weighting by live test levels to produce equivalent groups would be entirely circular, and the assumption that populations were equivalent across years would result in identical proportions at each level with no allowance for progress over time. NFER staff carry out live test to pre-test equating using the pre-test samples live test results, from the year preceding the one for which the test is being developed. This, according to NFER Staff, works reasonably well, but has to be adjusted by the pretest effect.

EdExcel staff do not look at differences between live test/pre-test and pre-test/pre-test equating, but the results are confounded by many factors.

OCR staff: No response.

Suggestion

No suggestions on this issue although some interesting ideas are discussed in the previous question.

Question 3

Some researchers argue that linking the results of one year to the results of the next may result to the accumulation of the error of measurement. Has any of the TDAs ever carried out an analysis to quantify/validate/reject this statement? Would it be more appropriate to directly link the data of one year back to the initial anchor data? Or is this paragraph totally not applicable/appropriate to mention in the context of the TDAs in England?

TDAs' responses with comment from the Consultant

NFER staff clearly suggested that linking to a baseline year has some merits, and could give interesting information if a careful study were conducted. However, NFER staff say, the TDAs are constrained by the brief supplied by NAA, which specifies linking to the preceding year.

EdExcel staff said that linking each test to the one immediately preceding it without reference to initial data clearly runs the risk of accumulation of errors; however, it is hard to see why errors would not average out at zero unless there was some systematic bias in the system, for example the new test was always administered after the old test; equated scores were always rounded down; the equating method was asymmetric, or some other design fault. EdExcel staff argued that they could equate to the original calibration of the anchor every year but this does not entirely remove the problem of accumulation of error. If the correlation between anchor test and pre-test scores decreases over time we may find that confidence intervals on equated values get progressively wider. In addition if there is a steady improvement in performance, matched by an increase in item difficulties, some software packages can start to have convergence problems for both items and pupils

OCR staff said that another strength of their equating model is that they equate back to the same year, rather than equating back to last year every time, thus avoiding the accumulation of error of measurement.

Consultant: The responses of the TDAs on this issue revealed interesting results:

(a) NAA may offer as an alternative to TDAs a link to the preceding year (for some good reason presumably).

(b) Even if TDAs linked to the original year, it is not certain that accumulation of error would be zeroed.

(c) It is has not been proved that linking to the preceding year indeed causes accumulation of error, and to what degree.

The NAA may like to explain why this is their suggestion. Recent suggestions by Massey *et al.* (2003), in agreement with the Consultant, include equating back to a baseline year

> *The current focus on year on year equivalence is an inherently weak strategy, in which the dangers of incremental drift in standards are readily apparent. Given medium-term curricular/assessment stability, we would recommend switching the focus of test equating, away from equivalence year to year, to a stepwise approach involving equivalence between a series of successive years and a 'stable' baseline before moving (at the transition between curricular cycles) cautiously to a new baseline, assuming that curricular changes then require it. This is the key to significant improvements in the quality of test equating possible, by comparison with current arrangements.* Massey *et al*. (2003, p. 232)

Suggestion

The responses of the TDAs show that this is a significant issue. However, it is not clear on which basis NAA '… specifies linking to the preceding year' (NFER staff). Does NAA think that this has an advantage towards linking to the initial year? Does NAA base this on a statistical basis which may be shared with all interested parties?

Further research on this would indicate whether accumulation of error is something TDAs should worry about. It would also evaluate NAAs brief to link to the preceding year. A carefully designed research proposal would shed much light on this issue. It would need to address the difference between the two methods, i.e. linking to the preceding year against linking to the baseline year.

Question 4

Have any of the TDAs made any study to specify minimum sample sizes for different datasets, for example subjects, types of questions? Do the current samples/methods used lead to equating errors that are smaller than half a mark (in the sense that one mark is the smallest measurement unit on which national curriculum levels are awarded)?

TDAs' responses with comment from the Consultant

NFER staff said that sample sizes are normally specified by NAA, although they sometimes ask for larger samples on the basis of experience and/or estimates of likely uncertainty. Their equipercentile equating produces confidence interval estimates, but when multiple 'chains' of equating are carried out it can be difficult to see exactly how to combine these measures of uncertainty. Confidence intervals of the order of +/- half a mark are normally achieved at scale mid-points with live test to pre-test equating using the whole main sample.

EdExcel staff said that the complexity of the key stage 3 test structure (four overlapping tiers) makes re-sampling/bootstrapping methods very time-consuming. EdExcel staff said that where it has been possible they have verified that equating errors are within acceptable limits.

OCR staff said that to achieve equating errors of less than half a mark, they would need a sample of at least 300 per test for English (which they achieve comfortably in their samples as they say) and a sample of at least 3,000 for science (which they fall short of, but are still comfortably, as they say, over the 350 needed to produce an error in the mean of less than 1.5 marks, which they feel is an appropriately small error, given the large mark range involved).

Suggestion

This question was more directly focused on the sample size and the error of equating, compared to the previous questions. The aim of the question was to directly investigate whether the TDAs were happy with the achieved precision of equating. Although they all appear at least satisfied, it is obvious that confidence intervals of half a mark are only sometimes achieved. QCA needs to decide if this is satisfactory, and then to take appropriate decisions if necessary.

In any case, the TDAs need to be consulted more directly before any decisions are taken.

Question 5

Have any simulation studies been used in this context to evaluate/quantify the effects/efficiency of different test equating designs?

TDAs' responses with comment from the Consultant

NFER staff said that simulation studies have not been carried out.

EdExcel staff say that they remain unconvinced of the need for this type of simulation and offer a number of arguments to support their case.

OCR staff: No response.

Issues not answered fully/clearly

The Consultant acknowledges that one of the TDAs' does not seem to be in favour of simulation studies in this context. This is fully appreciated, however, one of the TDAs gave no response at all to the question, and another one said that no simulations have been carried out. This might make one wonder whether this issue needs to be discussed more closely once again, just to clarify the position of the other two TDAs.

Software

This section deals with the software which is used by the TDAs to carry out their test equating tasks. The focus of this section is on the psychometric software, whether this is home made or commercial. The validity of the test equating results depends on the proper use of the most appropriate software for each task.

Different commercial statistical packages implement substantially different estimation algorithms, algorithmic shortcuts, modified versions of well known procedures etc. For example, in the Rasch 'world', Quest and Winsteps use the Joint Maximum Likelihood Estimation (Winsteps uses a modified PROX approximation at the initial stage as well), RUMM uses the PAIR method, ConQuest uses the Marginal Maximum Likelihood Estimation etc. They even use different fit statistics, different reporting conventions, and probably different rules of thumb to flag problems. The same issue probably happens in the IRT 'world', although the Consultant does not have specific evidence at the moment. This issue has already been addressed in the literature to some degree. For example, Pomplun *et al.* (2004) showed that Rasch packages can give different results in vertical equating.

The Consultant's personal experience with psychometric packages in general, makes him very cautious, and careful, especially when high stakes test equating is carried out. This is because there is no 'standard software' in the IRT world, as happens with SPSS and SAS, for example, in classical statistics.

Question 1

Are all the software packages licensed copies? Are they recently released versions of the software? Are all the relevant documentations, manuals etc. available to the users? Have all the users attended relevant training sessions on the use of the software, if and when necessary?

TDAs' responses with comment from the Consultant

NFER staff said that all software packages used by NFER are properly licensed, regularly updated versions and all manuals are made available to those who use each package. Training, either in-house or externally, is provided to those using the software package.

EdExcel staff said that they do not, and must not use unlicensed software and that they obtain updates as they become available. They attended training sessions where necessary.

OCR staff said that the relative simplicity of their equating methods enable them to carry out all the equating using Microsoft Excel and standard statistical programs such as SPSS and SAS.

Consultant: Two of the TDAs' use 'standard' commercial software, therefore no further research is needed on this. However, for the IRT software, personal experience, relevant published research and also informal discussions with colleagues make the consultant think that the TDAs should make sure that they have fully mastered all parameters and details of the software, since minor changes to the parameters may generate different equating results.

Suggestion

There may be differences in the results between different IRT software packages. This is not new; there are several published papers that deal with this. Whether this is a problem or not is up to anyone to decide. If different TDAs use different software, and if different software may give noticeably different results, then in the Consultants opinion, this is an issue.

It is not clear whether QCA should fund some comparability study on this, in the sense that there are commercial companies involved and possible legal issues may evolve. However, if a TDA uses some piece of IRT software, it is necessary to verify that the software works well on their data. Whether this could be done informally within the TDAs, or whether further research should be funded to this aim may be something to discuss.

Question 2

What formal procedures do the TDAs have in place to monitor the publication of literature relevant to the psychometric software they use?

TDAs' responses with comment from the Consultant

NFER staff said that there is no formal procedures in place specifically geared to monitoring literature of relevance to national test development. However, NFER's library has access to a wide range of international literature and alerts them to new books in relevant fields which come to their attention. At the same time, NFER's Chief Statistician, and other colleagues within the Statistics Research and Analysis Group, have as one of their professional goals the maintenance of awareness of current developments within relevant statistical literature, including psychometrics.

EdExcel staff said that a member of the Test Development Team has responsibility for cataloguing relevant papers.

OCR staff: Not applicable since no specialised software is used.

Suggestion

All significant updates released by the developers of the commercial packages used by the TDAs must be acquired and implemented. There needs to be a formal procedure by which they are updated about known bugs or other problems/news about the psychometric software used. Also academic papers demonstrating efficient use of psychometric software might be very useful.

Question 3

How and when were competitive commercial psychometric packages assessed to investigate the extent to which they are suitable for the specific datasets, sample sizes, types of tests etc. used in the national curriculum tests in the UK? For example, do competitive packages give the same results, and if not, which should we use?

TDAs' responses with comment from the Consultant

NFER staff said that from 2001 onwards the Chief Statistician has carried out formal evaluations of available psychometric (mainly IRT) packages, with the objective of deciding which should be the main tool for NFER to use in this area. NFER staff argued that as time passes, this area has to be re-visited regularly, as new software becomes available and the requirements for the packages change. They said that they have carried out some studies of results obtained by different packages, but these are hampered by different assumptions built into different models (e.g. Rasch versus 1-parameter IRT) and the ways in which estimation takes place.

EdExcel staff explained that they have tended to do this informally by obtaining evaluation copies to review if the software is user-friendly, reliable, consistent and fit for purpose. They have not done any direct comparisons, partly because the use of different fit statistics and reporting arrangements make direct comparison difficult.

OCR staff Not applicable.

Suggestion

Formal procedures are needed for the evaluation of new psychometric software. Both EdExcel and NFER staff acknowledge that there is a very poor consistency between various competing psychometric packages, and even comparisons of results are difficult to carry out. If one challenges the selection of specific software, particularly if the same data gives practically different results using a different package, there needs to be a short document to show why a particular package was selected.

Question 4

Have all the home-made software packages been properly developed, tested, verified and documented according to the ISO standards? Have they been evaluated, have their results been compared to similar home-made packages used by other TDAs? If they are not consistent, then which one should we use? Would it be possible/desirable for QCA, as an accountable regulator, to issue guidelines for this? What procedures were followed to answer the above questions and where is the evidence that this actually happened?

TDAs' responses with comment from the Consultant

NFER staff said that in-house software development tends to be mainly by means of macros written in a suitable language (for example, SPSS), and these are exhaustively tested and documented within the Statistics Research and Analysis Group, including benchmarks against existing methods, before being used on live projects. Documentation of these pieces of software is kept up to date, according to NFER staff, and files of problems, solutions and relevant warnings are also maintained.

Most importantly, NFER staff said that they would be happy (subject to resources) to test their software on other agencies' data, but would not be willing to share the source code with outside agencies because of concerns about intellectual property rights.

EdExcel staff said that they do not use home-made software packages, although they do have processes that use proprietary software adapted to their needs. The issue of ISO standards has not been raised previously and currently the Test Development Team at EdExcel staff have no experience in this area.

OCR staff Not applicable.

Suggestion

To the degree that home-made software is not used, this is not an issue. Sharing data for testing purposes between TDAs might help improving both software and research.

Documentation

This section deals with the documentation provided by the TDAs that carry out the test equating task. It is acknowledged by the Consultant that reports are very time consuming to produce and experts' time is too valuable for the TDAs to 'waste' on report writing. However, documentation is needed in order to be able to answer questions set by the public or other interested parties about past test equating procedures. In the case where the thoroughness of the procedures followed are under question, detailed reports may give the opportunity for QCA and the TDAs to stand up and defend their results.

The literature has already dealt, to some degree, with the issue of reporting the output of statistical analysis, and especially of the Rasch/IRT statistical analysis. For example, Lamprianou (2006a) has studied how Rasch analysis results have been presented in a specific Rasch-specialising Journal for a number of years. Smith, Linacre and Smith (2003) also issued specific guidelines on how to report the results of Rasch analysis. Similar guidelines might be used for the presentation of the IRT results. For example, Rupp and Zumbo (2004) suggested how to quantify and report whether IRT parameter invariance holds and studies the case when other statistics like Pearson correlations are not enough.

Question 1

Have we ensured that our reports describe the procedures, methods and software in such detail so that an independent researcher can replicate or question our findings?

TDAs' responses with comment from the Consultant

NFER staff: No response.

EdExcel staff want their reports to be informative and fit for purpose and welcome clarification on what level of detail is needed. They are not operating as an academic establishment and they should not be expected to reveal specialist knowledge that might be prejudicial to their commercial interests.

OCR staff said that the documentation they produce for the Draft Level Setting (DLS) report is generally quite brief and that although it would be a little more work to produce an additional Technical Report to go alongside the Draft Level Setting report, it would be useful to have at the DLS meeting, so that specific queries relating to the details of the equating could be answered immediately, if raised.

Consultant: The Consultant acknowledges that the TDAs should not be expected to reveal commercially useful knowledge. However, giving enough information in order to replicate the equating in the future if needed, seems to be important, in order to protect the interests of QCA, and the stakeholders as a whole.

Suggestion

QCA needs to take a decision on this issue. It affects transparency, quality checks and reproducibility of results. The Consultant agrees fully with Massey *et al*. (2003) who suggested that external audits of the test equating procedures and results will reassure the public about the consistency of standards across time:

> *…it may provide some public re-assurance if it was understood that, after an extended period, independent audits of threshold setting decisions in each key stage/curriculum area were to be undertaken and their reports made public. Audits would consider equating evidence…threshold recommendations and distributional data, and the other sources of information available….Files which would be suitable for audit at a later date would need to be produced annually, consisting largely of… papers, with some additional details regarding equatings. It might be desirable to ask an auditing agency to observe the standard setting process each year and to participate in the construction of the files, to ensure that these prove fit for purpose...* Massey *et al. (*p.237, 2003)

The Consultant is not in a position to know whether this recommendation was discussed in the past and was rejected by QCA. In any case, external audits are always welcome and they would enhance the trust that the public puts on the equating results.

Question 2

Which section of the report is aimed at the technically literate audience, and which part of the report is aimed at the general public? What is the content?

TDAs' responses with comment from the Consultant

NFER staff: No response.

EdExcel staff said that the draft level setting report is not aimed at the general public and that they need to decide on the appropriate level of detail for the intended audience rather than try to cater for the knowledge level of all potential interested parties.

OCR staff: No response.

Suggestion

QCA needs to decide whether there is a need for a more and a less technical part of the report. This question needs to be put in the context of the previous one, and especially in the context of the suggestion for external audits.

Question 3

In the case of the detailed 'technical report' (which, perhaps, could be released at a later stage if time is an issue), how do we explain/describe the following:

a. *Argument for the selection of the specific software.*
b. *Argument for the specific statistical model.*
c. *Data cleaning and verification.*
d. *The representativeness of the sample.*
e. *The quality of the sample (in addition to the representativeness).*
f. *The 'program files' used (for example, most of the psychometric software needs little 'control files' which give all the commands to the software in order to run the analysis and produce the output). If these are missing, how can a researcher replicate our findings?*
g. *The 'stopping rules' need to be explicitly specified (for example, if index X is that large we cannot continue the analysis because the model-data fit is not acceptable).*
h. *Contingency plans to be triggered in the case of 'stopping rules' actually reached, for example additional sample if the error of measurement is large; or, what happens if the model-data fit is not 'acceptable' based on pre-defined criteria?*
i. *The raw output of the analysis.*
j. *The actual sample used.*
k. *Our interpretation of the results (qualitative judgments like 'satisfactory model fit' need to be justified on specific criteria). Specially referenced should be: error of measurement, mode-data fit.*
l. *What special tests have been done to verify that the properties of the IRT models (when we use IRT) hold? For example, if the item-invariance does not hold to a degree, then is the equating procedure in danger?*
m. *Formal suggestion for the results of the test equating.*

n. *Advice for the level setting procedure: for example, the statistician may think that this year the statistical equating may be biased or inaccurate, therefore, other evidence like test scrutiny might play a more significant role than usual this time.*

TDAs' responses with comment from the Consultant

NFER staff: No response.

EdExcel staff gave a response to almost all sub-questions. However, in some cases the response was too brief.

OCR staff: No response.

Consultant: EdExcel staff attempted to give responses to all sub-questions. Maybe this topic needs to be re-discussed in the future, if the TDAs also think that this is an important question.

Discussion and recommendations

Both the independent Consultant and the TDAs have invested a lot of time and effort on this report. The whole procedure took many months, and the TDAs provided responses to many questions. In some cases the TDAs did not provide responses to the questions and this is clearly indicated on a question by question basis. Sometimes the questions were not applicable, for example some questions were almost exclusively focusing on IRT issues; in such cases, the TDAs that do not use IRT could skip the questions. However, in certain cases some TDAs simply chose not to respond. For example, in the section under the title 'The quality of the datasets/samples' question 3 was asking a very straightforward question: *Test equating needs to comply to the relevant 'Code of practice'. How is compliance with the 'Code of practice' monitored by the TDAs themselves?* Still, one TDA chose not to give a response to this straightforward question.

Although the TDAs skipped a number of questions, generally speaking, the responses of the TDAs shed plenty of light on important issues. This document is very lengthy and needs a very careful reading; the messages may be too dense sometimes. This conclusion does not attempt to summarise the whole document; rather, the interested reader is strongly encouraged to go through the whole report.

During the technical seminar that was held in 2006 the Consultant had the opportunity to talk to members of the TDAs, attend their presentations and ask clarifications about the methods they use, how they use them etc. The TDAs were very open, and it was obvious that their analysts were very confident, experienced and applied the test equating models in a very professional way, comparable to how the models are used in similar contexts across the world.

The task of supporting the standards across time is a collaborative one. The Consultant does not think that QCA should over-interpret the results of this study by taking very important unilateral decisions. Decisions should be taken under the light of further discussions and negotiations with the TDAs.

Before setting up another technical conference, however, QCA first needs to answer some questions, in order to be prepared to take decisions. There are several major

messages from this lengthy document that actually signal questions to QCA. For example:

(a) Is it desirable (politically and/or administratively or otherwise) for QCA to encourage the TDAs to use the same test equating methods? If yes, should QCA enforce uniformity by means of relevant articles in the Code of Practice or should research results be used to convince the TDAs to move towards uniformity (i.e. reach consensus)?

(b) Should the error of equating be the same for all subjects, or are there any reasons to be more lenient for science and more demanding for mathematics (other than the cost of the sample size)? Currently, the error of equating for science tests is several times larger than the error of equating in mathematics. QCA needs to take a decision on this issue, taking into account practicalities as well as costs, and then to discuss the decision with the TDAs.

(c) The equating errors are analogous (among other things) to the sample size used for the test equating exercise. The question is *how are the sample sizes for each test equating exercise specified and who specifies them*? It has been suggested by one respondent that the sample size is decided by the NAA. We need to know, however, how the estimation of the sample size is done, and the arguments behind it. QCA needs to explicitly declare how the estimation of the sample size needed should be carried out. For example, using a backwards strategy, the sample size may be computed directly from the desired equating error. Following a forward strategy, the sample size will be computed depending on the cost or the availability of the schools (having said that, as a Consultant, I do not personally think that the cost should be the major criterion in this case).

(d) Is it politically desirable to enforce (or indeed negotiate) some sort of sporadic and external audit of the test equating procedure/results, as described in this document? This could increase the public trust; however, one might want to consider the increased cost.

(e) Related to the above is the need for better (more lengthy and possibly more thorough) technical documentation. Again, there are direct cost implications; would QCA be happy to see a rise in costs?

(f) Finally, probably the most important question, would QCA be happy to support relevant test equating research in the context of the national curriculum tests? Although there are many international publications about test equating, there is not much published knowledge on test equating in the context of the English national curriculum tests. If QCA would like to pursue further research, then a technical conference could define research areas, and decide who is going to undertake the research and who is going to provide the budget.

These are the Consultant's recommendations:

- Statistical models: A carefully designed research (in coordination and cooperation with the TDAs) is certainly needed in order to investigate whether the 'competing' models do indeed give noticeably different results. If different models give substantially (practically and statistically) different results, we need to know why we use the models we currently use.
- Sample sizes: QCA needs to decide (and reason) which error margin is acceptable. If an error of 1.5 or 2 marks at the mean of the scale (presumably much larger at the tails) is acceptable, it is important to investigate what % of students might have been awarded a different level, had the threshold been 1.5 or 2 marks up or down. The same error margins should exist for different subjects or different key stages.
- Documentation: It should be enriched with additional information, in order to improve both transparency and reproducibility of results. This might facilitate/feed random and sporadic external audits to reassure the public about the quality of the procedures and the equating results.
- Transparency and accountability: QCA might want to investigate ways to help TDAs publish their draft research findings. For example, an on-line library with technical reports, with adequate information and available datasets to replicate the analyses might be a very good research resource in order to guide decision taking.
- Pre-test effect: The investigation of the so called pre-test effect is both academically and practically significant and needs to be investigated more systematically.

- Linking of standards to previous year: NAA apparently specifies linking of new test results to those of the preceding year instead to those of a baseline year (where possible). The TDAs, in their responses, do not deny that this design might be susceptible to accumulation of error. Further research on this would indicate whether accumulation of error is something the TDAs should worry about.
- Unilateral decisions should not be taken by QCA, without discussing issues of costs and practicalities with the TDAs.

References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed., pp.508-600). Washington, D. C.: American Council on Education (Repreinted as W. A. Angoff, Scales, norms and equivalent scores. Princeton, N.J.: Educational Testing Service, 1984).

Athanasou, A. J., & Lamprianou, I. (2002). A teachers' guide to assessment. Sydney: Social Science Press.

Baker, J. G., Rounds, J. B., & Zevon, M. A. (2000). A comparison of Graded Response and Rasch Partial Credit models with subjective well-being. Journal of Educational and Behavioural Statistics, 25 (3), 253-270.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. Applied Measurement in Education, 12 (4), 383-407.

Bramley, T. (2006). Equating methods used in KS3 Science and English. Paper for the NAA technical seminar, Oxford, 23-24 March 2006.

Budescu, D. V. (1987). Selecting an equating method: Linear or equipercentile? Journal of Educational Statistics, 12 (1), 33-43.

Demars, C. E. (2005). Type I error rates for Parscale's fit index. Educational and Psycological Measurement, 65 (1), 45-50.

Ferrara, S. F., Huynh, H., & Baghi, H. (1992). Assessing local dependency in examinations with clustered free-response items. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Green, D. R., & Langhorst, B. H. (1986). Passage dependency and item characteristics. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Hambleton, R. K., & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Boston: Kluer-Nijhoff Publishing.

Hanna, G. S., & Oaster, T. R. (1980). Studies of the seriousness of three threats to passage dependence. Educational and Psychological Mesurement, 44, 583-596.

Hanso, B., Zeng, L., & Colton, D. (1994). A comparison of presmoothing and postsmoothing methods in equipercentile equating. ACT Research Report Series, 94 - 4, ACT.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. Applied Measurement in Education, 6, 195-240.

Hayes, M., & Field, R. (2006). EdExcel Test Development Team: Equating methods in National Curriculum Assessment of Mathematics (KS2 and KS3). Paper prepared as background to the NAA seminar on equating methods used in national curriculum tests, Oxford, 23-24 March 2006

Holland, P. W., & Rubin, D. B. (1982). Test equating, New York: Academic.

Huynh, H., & Ferrara, S. (1994). A comparison of equal percentile and partial credit equatings for performance-based assessments composed of free-response items. Journal of Education Measurement, 31 (2), 125-141.

Kolen, J. M. (1981). Comparison of traditional and Item Response Theory methods for equating tests. Journal of Educational Measurement, 18 (1), 1-12.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. Journal of Educational Statistics, 9 (1), 25-44.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. Journal of Educational measurement, 28 (3), pp. 257-282

Kolen, M. J., & Brennan, R. L. (2004). Test equating, Scaling and Linking. Methods and Practices. Second edition. New York: Springer-Verlag.

Lamprianou, I. (2006). The stability of marker characteristics across tests of the same subject and across subjects. Journal of Applied Measurement, 7 (2), 192-200.

Lamprianou, I. (2006a). Development of a 'Comprehensiveness of Rasch Measurement Application' scale. Journal of Applied Measurement, 7 (1), 92-116.

Lamprianou, I., & Boyle, B. (2004). Accuracy of measurement in the context of mathematics National Curriculum tests in England for Ethnic Minority pupils and pupils who speak English as an additional language. Journal of Educational Measurement, 41 (3), 239-260.

Levine, R. S. (1955). Equating the Score Scales of Alternative Forms Administered to Samples of Different Ability (RB-55-23). Princeton, N.J.: Educational Testing Service.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. Journal of Educational Measurement, 30 (1), pp. 23-39

Livingstone, S. A. (1993). Small-sample equating with log-linear smoothing. Journal of Educational Measurement, 30 (1), 23-39.

Lord, F. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Massey, A., Green, S., Dexter, T. & Hamnett, L. (2003). Comparability of national tests over time: Key stage test standards between 1996 and 2001. Technical Report, Qualifications and Curriculum Authority, London, UK.

NfER (2006). Equating Methods in National Curriculum Assessment . Presentation handounts prepared by NfER for the QCA Equating Seminar 23-24/3/06, Oxford.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp.221-262). New York: Macmillan.

Pomplun, M., Omar, H., & Custer, M. (2004). A comparison of Winsteps and Bilog-MG for vertical scaling with the Rasch model. Educational and Psychological Measurement, 64 (4), 600-616.

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: when pearson correlations are not enough. Educational and Psychological Measurment, 65 (4), 588-599.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. Educational and Psychological Measurment, 66 (1), 63-84

Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. Journal of Educational and Behavioural Statistics, 25 (4), 391-415.

Smith, R. M, Linacre, J. M., Smith, E. V. Jr. (2003). Guidelines for manuscripts. Journal of Applied Measurement, 4 (2), 198-204.

Tate, R. L. (2003). Equating for long-term scale maintenance of mixed format tests containing multiple-choice and constructed response items. Educational and Psychological Measurement, 63 (6), 893-914.

Wang, W., Chen, C. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. Educational and Psychological Measurement, 65 (3), 376-404.

Wright B. D. (1967). Sample-free Test Calibration and Person Measurement. ETS Invitational Conference on Testing Problems. Chicago: Mesa Psychometric Laboratory. The document was reached at http://rasch.org/memo1.htm#top at 12/04/98

Yen, W. (1992). Scaling Performance Assessments: Strategies for Managing Local Item Dependence. Invited address at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

