

National Foundation for Educational Research

Technical seminar on

**Statistical Methods used for the
Analysis of National Monitoring
Surveys**

Thursday 28 January 2010, at The National Foundation for Educational Research

Proceedings

Edited by Marian Sainsbury, Jo Morrison and Sarah Maughan

Contents

Introduction	3
National Monitoring: the Context	4
Summary of presentation.....	4
Session 1 Sampling.....	5
Summary of presentations	5
Discussion on Sampling	7
Session 2 Links to TIMSS	8
Summary of presentations	8
Discussion.....	9
Session 3 Analysis	11
Summary of presentations	11
General Discussion.....	14
Concluding Remarks.....	15
Appendix 1: List of participants	17
Appendix 2: National Monitoring: the context	19
Appendix 3: Statistical Methods used for the Analysis of National Monitoring Surveys - Sampling	22
Appendix 4: Recovering information in the NPD using multiple imputation	34
Appendix 5: National Monitoring: Link to TIMSS?	38
Appendix 6: TIMSS – working with what we already have	42
Appendix 7: Item Response Theory.....	45
Appendix 8: Domain sampling and generalizability theory	50

Introduction

Since October 2008, national assessment policy in England has been in a state of change. At that time, the Secretary of State for Children, Schools and Families announced the end of national testing at Key Stage 3, a system that had been in place since the mid-1990s. This announcement was followed by the deliberations of an Expert Group to consider revised arrangements, which reported in February 2009, a report to which the Secretary of State gave a formal response soon after.

A central element of the new assessment landscape is the intention to establish a national monitoring survey. Originally, the introduction of such a survey had the intention of monitoring standards only at the end of Key Stage 3, where there were no longer national tests. Later, a national monitoring survey for science at the end of Key Stage 2 also became policy, in response to the Expert Group recommendation of discontinuing the science tests at this key stage. The group further recommended that national monitoring should be linked to existing international surveys.

Previous expert seminars held by the National Foundation for Educational Research (NFER) during 2009 considered some of the implications of these changes. In January 2009, researchers and policy-makers discussed national assessment arrangements for Key Stage 3 at a seminar hosted jointly by NFER, Cambridge Assessment and The Nuffield Foundation (Sainsbury and Maughan, 2009). In June, 2009, NFER and the Chartered Institute of Educational Assessors held a seminar on methods for ensuring reliability of teacher assessments in the new context (Parkes and Maughan, 2009).

The complexities of planning a national monitoring survey were introduced at the January seminar, and it was clear at this point that there were many detailed technical issues in need of examination. This latest seminar was set up by NFER to respond to that need by stimulating discussion about the different statistical techniques that might be useful for the proposed national monitoring surveys.

In particular the seminar sought to question:

- issues and solutions involved in sampling
- possibilities for links to the TIMSS international survey
- statistical techniques available for analysis.

Contributions were made by Simon Rutt, Dougal Hutchison, Graham Ruddock, Ben Styles and Tom Benton (NFER), Harvey Goldstein (University of Bristol) and Sandra Johnson (Assessment-Europe). The seminar was steered by Sarah Maughan and Chris Whetton (NFER).

This report summarises the presentations and discussions. The seminar operated under the Chatham House rule, but the presenters have given permission for their views to be attributed. A list of attendees is included as Appendix 1.

National Monitoring: the Context

Sarah Maughan, NFER

Summary of presentation

This presentation appears in full as Appendix 2.

Sarah provided a brief introduction to the day by describing the preparations for the introduction of a national monitoring survey in England. After the abolition of the Key Stage 3 tests in October 2008, the Expert Group on Assessment (NFER, 2009) recommended that a *'sample testing system should be introduced for pupils at the end of year 9'*. The Expert Group also recommended that Key Stage 2 science tests should be abolished. When accepting the recommendations of the group, the Government also introduced the idea of a sample testing system to replace the Key Stage 2 science tests.

The current expectation is that national monitoring will be introduced in its final form in 2012. For Key Stage 2 science, the existing tests will be used in 2010 and 2011 with a sample of pupils; for the Key Stage 3 tests a pilot of a new model of assessment will be introduced in 2011.

There is a long history of sample testing that we can learn lessons from: the Assessment of Performance Unit in England in the early 1980s; national monitoring in other countries including the USA, Scotland and New Zealand; and the international surveys: PIRLS (Progress in International Reading Literacy Study) , TIMSS (Trends in International Maths and Science Study) and PISA (Programme for International Student Assessment).

Sarah posed a number of key questions that require answers before progress can be made: what is the purpose of the sample testing, what model should the testing follow and what reporting requirements will there be? However, the question to be focused on in the seminar is: what are the technical considerations for the sampling and analysis aspects of the new sample tests?

Session 1

Sampling

Summary of presentations

Simon Rutt and Dougal Hutchison, NFER

Harvey Goldstein, University of Bristol

These presentations appear in full as Appendices 3 and 4.

Simon Rutt's session on sampling for a national monitoring exercise started with reference to the opportunity that exists for its creators in restoring some trust and confidence in the assessment system. There has been a large amount of criticism of the current assessment practices from the teacher unions to academics and parents. Any new system must stand up to the close inspection and criticism that will inevitably come its way. If decisions are based on sound methodological and statistical principles then rational criticism can always be defended with a rational design. It is vitally important that the purpose of the system is agreed before designing the system. What exactly is the aim of the assessment and what is the best design to achieve that aim?

There are a number of important questions that need to be answered when designing the sample and some of those questions could include the following:

- Is the monitoring system a single analysis of national performance or is it to be combined with analysis of sub groups? What might we want in the future: analysis by gender, SEN, ethnicities, deprivation, vulnerable groups?
- Will the system require analysis at item level or groups of items linked to assessment focus areas? For example will the system assess all areas of the Key Stage 3 mathematics curriculum?
- What stratifiers would be needed, e.g. school, pupil or prior attainment?
- At what level are pupils selected – whole year group, whole class, within class? A pupil sample could use the National Pupil Database to select pupils from within classes but the effect on individual pupils selected would need to be thought through. Selection is anonymous at the national level but everyone within a school would know who had been selected. Selecting at class level results in a degree of uncertainty in knowing who is being selected.

The design of the sample therefore covers many considerations but simply the sample design will be affected by what is the assessment for, who is it for, what level of precision is required, what budget is available and what level of burden on teachers and schools is right for the project.

Dougal Hutchison explained that while a national assessment looks like ordinary testing, it actually differs in important aspects. From the point of view of design and analysis of the study, the major challenge is that the number of test items involved is generally too large for any single individual pupil to sit, so a design is required in which all items are taken, but not necessarily by the same pupils. He described how the survey can be designed to link the components of such a test design, using a cyclic or 'cartwheel' design. From the point of view of pupils, schools or teachers, this means that it is low stakes, since it does not directly produce comparable individual scores, and results do not affect the individuals involved.

A sample of 400 pupils sampled at random from the entire country (a 'simple random sample') could be sufficient for very simple results, but the sample size has to be increased because of the design of the test to cover a very large number of items, with at least 400 pupils taking each version. Further, there is generally some kind of clustering of pupils sampled within schools, and while sampling via schools is more administratively convenient and economically efficient, school administrative policies such as setting mean that less independent information is gathered, and larger samples are again required. Dougal presented some graphs to show how this can affect the sample size required, and how much precision can be obtained from a given design.

Harvey Goldstein explored the implications of only testing a sample of pupils on the subsequent value-added analysis of the National Pupil Database, by considering it as a missing data problem. Investigation focused on conducting a value-added analysis between KS2 to KS4, as there is a high correlation between these two key stages and KS4 data would have complete coverage. Data that is missing by design can be analysed using multiple imputation methods, meaning that secondary analysis, for example regression and hypothesis testing, would be possible at a national level.

A 2-level model with covariates gender, eligibility for free school meals and KS2 result was fitted on the full dataset. A number of different samples, each comprising 15,000 pupils (3% of the cohort), were taken using different sampling strategies. Missing KS2 values were imputed using multiple imputation techniques, and the subsequent models were compared to the full model.

The sampling strategies were to randomly select:

- 3% of pupils across schools
- 17% of pupils within 50 schools
- as above but to add ethnic group and SEN status into the imputation model

The results showed that with efficient sampling and use of multiple imputation methods secondary analysis was useful, with multiple imputation methods reducing the standard errors, and with the inclusion of other background variables resulting in further improvements. If pupils were sampled across schools, not within schools, this strategy would not be compatible with league tables, as the school level standard errors are too large because of the small number of pupils sampled in each school.

This work was an initial exploration of the possibilities and further investigation of complex sampling design and use of other covariates (at the school and pupil level) would be necessary, for example possibly building in connections with pupils at KS3, or rotating background variables, or including compositional effects if most children in a school were sampled.

The technique was computer intensive and highly statistical.

Discussion on Sampling

The discussion mainly formed two strands; technical aspects and purpose of testing.

The technique described by Harvey Goldstein would be equally applicable to subjects other than, and potentially less reliable than, maths, which was used in his exploration. Categorical outcomes or predictors could be incorporated as well as continuous variables. It would be possible to include independent schools in the analysis as many secondary independent school pupils were in state primary schools. Although the relationship between KS2 and KS4 may change over time, this would not impact, since for each year the analysis would utilise the current relationship, and the purpose is not to make future predictions. The National Pupil Database (NPD) would be used to calculate sample sizes. The cost of computer power and analysis time would be offset by savings in testing a sample of pupils rather than nationally.

The purpose dictates the sample size and design. If desired the design could include a time element, or links across key stages. Incorporating a matrix testing design would impact on sample sizes, as would any desire to do analysis of subgroups, e.g. sex differences. Sampling different schools in different years can be incorporated if there is an interest in school performance over time. The existing NPD could be used to calculate statistical power, appropriate for the chosen sample design.

Session 2

Links to TIMSS

Summary of presentations

Graham Ruddock, NFER

Ben Styles, NFER

These presentations appear in full as Appendices 5 and 6.

Graham Ruddock's presentation addressed the analysis of curriculum match that took place as part of the TIMSS surveys. In both the 2007 and 2003 TIMSS surveys, curriculum experts from England reported that in mathematics over 95% of the TIMSS items were judged to be in the curriculum in England for both grades, Grade 4 (Year 5) and Grade 8 (Year 9). For Grade 8 science the figure was slightly lower, with grade 4 science given the lowest ratings, 81% in 2003 and 69% in 2007. The lower ratings for science are caused, in part, by the inclusion of earth science in TIMSS, which does not appear in the science curriculum in England.

Further work on the match between TIMSS science and the national curriculum in England has been funded by NFER. This confirmed the strong relationships between the two except in earth science. These two sets of judgments, taken together, point to the suitability of using TIMSS as part of a national monitoring exercise in England.

The presentation also looked at issues and possibilities in linking the findings from a national survey to the international results. One issue raised was whether to use only those TIMSS items being held secure at the time of a national monitoring exercise, just over 50% of TIMSS items, or to include released items made available to the public domain. Possibilities for linking to TIMSS in Year 9 include linking directly to the TIMSS IRT scale and/or linking to the international benchmarks defined at set points on the TIMSS scale. It would also be feasible to link monitoring in Year 6 (Grade 5) to the TIMSS Grade 4 (Year 5) scale. A further possibility is offered by the recent model of national curriculum tests being longer than the TIMSS tests. This would allow a 'TIMSS plus' approach, comprising all of TIMSS plus a range of England only items, to be administered without increasing assessment time for students or schools. Attitudes to mathematics and science could also be monitored using questions from the TIMSS student questionnaires.

Ben Styles presented work that he had done with Naomi Rowe, entitled 'TIMSS – working with what we already have'. The presentation summarised a methodology for transforming relative strengths and weaknesses in grade 8 science, as highlighted by TIMSS 2003 and 2007 studies, into the 'language' and context of the English Key Stage 3 science curriculum. The approach involved mapping trend items, from TIMSS 2003 and 2007, to domains of the English national curriculum (completed). The output from this method of grouping items will

be mean percentages of items correct in TIMSS 2003 and 2007, in order to show trends over time (work still to be completed).

Trend items were grouped into England's national curriculum 'domains': Sc1 Scientific Enquiry, Sc2 Biology, Sc3 Chemistry and Sc4 Physics. The resulting mean percentages and their corresponding standard errors will be calculated as per IEA methodology (Beaton and Gonzalez, 1997). This involves calculating the mean percentage correct across selected items using the overall sampling weight. Each replicate weight is then used to calculate a mean percentage correct. The variation between the original sample estimate and the estimates from each of the replicate samples is the jackknife estimate of the sampling error of the statistic. Any changes between 2003 and 2007 will be tested for statistical significance. Mean percentage correct will also be computed for all science trend items as matched to England's curriculum in order to provide a point of reference for specific domain trends. A key element in this analysis is rather than attainment being presented in the context of the TIMSS assessment framework for science, attainment will be given in the format of England's national curriculum 'domains'.

Discussion

The discussion was wide-ranging, encompassing both theoretical and practical questions.

Purpose of linking national and international surveys

The fundamental question was raised of why it might be desirable to link national and international surveys at all. Two possible justifications were given. International data gives an external check on the findings of national assessment. This brings an outside perspective which can cast light on the controversies that often surround claims and counter-claims about national findings. Secondly, the link to international surveys was an explicit recommendation of the Expert Group, which gives it a high priority in policy making.

Reporting issues

It was agreed that the introduction of this new system should signal a switch to a new national standard. It would not be possible to maintain continuity of standards from the previous national curriculum tests to the new system, particularly as this would be a move from a high-stakes to a low-stakes assessment. It should be established and communicated that 'In the year of change there is no change' – that is, that standards should be regarded as unchanging as the measure changes from old to new.

Further, it should be established that, as a sample survey, any results would have the status of estimates, presented with confidence limits. This would be a departure from national curriculum test results, which are population measures rather than sample estimates.

Most international surveys have a reporting scale with a wide score range, often with a mean of 500. This differs from the level-based reporting with which the public in England is familiar, and which is perceived as easy to interpret. A solution might be to have both, in a

similar way to PISA, which has levels with descriptors. At a national level, it must be possible to calculate what proportion of pupils achieve a level and interpret this in terms of skills. This involves a transformation from scale score to descriptor, normally a judgmental process involving a group consensus. The judgmental, rather than objective, nature of this process needs to be carefully communicated to the public.

Frequency of surveys

It was agreed that an annual survey may be too frequent, for two main reasons. Firstly, a year is too short a period to detect any significant change. Secondly, there is no time to draw lessons from the findings of one survey before the next is administered. One recommendation might be a planned cycle of subjects so that, whilst the survey might have a one-year cycle, each subject assessed might appear only every three years, allowing time for learning and adjustment between each one.

TIMSS issues

It was agreed that, if TIMSS were to be used as part of national monitoring, it should be used in its entirety, rather than extracting parts. Participants noted the experience of the US National Assessment of Educational Progress (NAEP), where attempts had been made to adapt TIMSS, resulting in questionable analyses and difficulty of interpretation.

The TIMSS items belong to the International Association for the Assessment of Educational Achievement (IEA), and permission to use them would have to be negotiated. It was agreed that IEA should be asked to approve the specific proposed design for using TIMSS as part of a national assessment for England, as a check on its suitability.

There was also some discussion about the security of TIMSS items. As Graham's presentation made clear, IEA keep half of the TIMSS items secure at any one time, and it would be necessary for this security to be maintained when they were used in the national monitoring survey.

Session 3

Analysis

Summary of presentations

Item Response Theory (IRT)

Tom Benton, NFER

This presentation appears in full as Appendix 7.

Item response theory is about modeling all of the influences on the chances of a student getting a particular item correct. Its traditional application, such as in the major international studies, focuses on the relationship between a single underlying measure of ability and this probability but recent advances in both theory and software mean that this approach need not be followed. It is now quite possible to use extended IRT models to take account of the different behaviour of items for different subgroups or of the particular relationships between different items.

Item response theory is a useful tool in that it allows us to link scores across different tests. This allows us to use many items without any individual pupil having to sit an overly long test. IRT is useful in test design in that once we have built our model we are able to select items such that our final tests have the desired characteristics in terms of difficulty and reliability. It is also useful in reporting in that we can estimate changes in standards over time alongside properly calculated confidence intervals for these estimates as well as being able to give results to students on a common scale if this is desired.

There are many different possible IRT models ranging from the relatively simple to the highly complex. Each of these models makes different assumptions and it will not always be obvious which is the most appropriate for a given set of data. As such it may be worth considering sensitivity analysis as a part of the way these methods are applied. If two equally valid models give very different results this may lead to difficulties in definitively interpreting results. The extent of difference between equally valid models may provide a sensible limit on how accurate we should attempt to be within a sampling framework. A brief analysis of TIMSS grade 4 maths data revealed little to be worried about in terms of differences between alternative IRT models although this may reflect something about the way these tests are constructed to fit within a particular IRT framework.

Domain Sampling and Generalizability Theory

Sandra Johnson, Assessment-Europe

This presentation appears in full as Appendix 8.

The introduction of a sample based survey approach to replace national testing at Key Stage 3 will presumably have the principal aims of providing population estimates of pupil performance in the main subject areas within and over time, along with estimates of the performance of subpopulations (e.g. boys and girls). It would be desirable for the surveys also to gather information on teaching and learning to provide a background against which to interpret pupil performance.

Pupils will be sampled from within the relevant pupil population, and since pupils are clustered with schools and have different background characteristics a complex sampling scheme will be employed. The subject domains to be assessed will essentially be the Key Stage 3 curricula in the relevant subjects, each of which covers a range of subject knowledge and skills. For the purposes of assessment the subject domain has to be operationalised through the creation of valid and useable test questions (items). Subject domains can rarely be comprehensively operationalised, however, since it is rarely possible to have available all the possible test items that would jointly represent it. In practice we develop as many appropriate items as we can, to create as large and as representative a question pool as possible. From within this representative pool items are sampled for use in surveys in the same way that pupils are sampled for testing.

Sampled items are distributed among sampled pupils using a matrix design, and pupils' performances on the items are analysed in an appropriate way to produce domain-referenced population attainment estimates and margins of error. The underlying measurement model should allow for the presence and influence of numerous different sources of measurement error. These will in particular include interaction effects of various kinds: for example, between markers and items, between pupils and items, and between pupil subgroups and items ("differential functioning", a phenomenon which cannot necessarily be considered as non-valid, and which should not automatically be addressed through item rejection). The principal strengths of the domain sampling approach were identified as validity, simplicity, transparency and comprehensibility.

Discussion

The following summarises the discussion which centred on two main themes.

Domain Sampling

It was acknowledged that domain sampling ideally requires large item banks to function well, and that it might be more readily applied in some subjects, e.g maths and science, than others. It would be applicable for reading tests by constructing short batteries of items relating to one short passage, analysing using testlet theory.

Statistical models

Model choice is important statistically, philosophically and from the perspective of interpretation. It was acknowledged that statistical models including IRT, multilevel modelling and generalisability theory had improved considerably since earlier years, and continue to become increasingly sophisticated. All allow interactions to be incorporated and it was important to include these to allow for differences between different types of pupils and schools. It was felt important to include contextual data, since policy makers really want to know what drives changes in attainment. Any items showing unwanted characteristics e.g. dependence or differential item functioning, need to be examined and not automatically discarded, although attempts to explain or justify DIF, for example, are often no better than random. Models should not be constrained by unidimensionality. Multidimensionality should be embraced and explored. However examining changes over time requires enforcing some bold assumptions so it is not desirable to include items that violate these. And some of these assumptions are separate from the choice of statistical model, they are philosophical; are changes over time really changes in pupils' ability? Different models make different assumptions: IRT assumes that item difficulty is stable, including over time, while domain sampling depends for over-time interpretation on the assumption that the nature of the domain is unchanged, or that any change is known and its effects on pupil performance predictable.

General Discussion

Many broader points were made in the final discussion, picking up aspects of different presentations in the course of the day.

Purpose and outcomes of national monitoring

The discussion returned to the importance of defining the purpose of a national monitoring survey and wider considerations of its feasibility. Several possible perspectives were put forward. The survey might aim to monitor standards over time. Alternatively, it could be seen as providing evidence about what conditions lead to high attainment. Or, again, it could be a means of researching the relationship between curriculum and attainment.

There was a discussion of whether it is actually possible to monitor standards over time, in the face of changes to the educational and social environment and to curriculum and testing. The overall view seemed to be a qualified 'yes'. Standards can be monitored over the short term, but not reliably over more than about ten years. There may be more stability in examining relationships, rather than simple scores, over time – for example, the relationship between boys' and girls' scores, or between outcomes in different parts of the country. The survey offers the possibility of documenting changes in how skills are defined.

Beyond this, there are significant issues about how the results are interpreted. There was a view that stable results are not desirable for policy-makers, who prefer to be able to report changes. To guard against distortions, it is essential that a monitoring survey retains low stakes.

There was also some discussion of how 'Key Stage 3' might be defined, in the absence of the national tests. Some schools are now regarding Key Stage 3 as a two-year stage, starting Key Stage 4 in year 9. In these cases, national monitoring of year 9 pupils could lead to depressed results because content covered in Key Stage 3 was no longer immediate and had been forgotten.

Combining the strengths of different approaches to analysis

The positive potential of recent developments in statistical analysis was discussed. The latest versions of analysis programs had a power and flexibility that was lacking in previous years. This holds out the possibility of a more sophisticated approach to analysis that builds on the strengths of all the approaches discussed.

It was agreed that an integrated framework, drawing on domain sampling as well as IRT, could be a particular strength of a new system. However, this combination of analyses should be approached cautiously. Different possibilities should be researched and reported transparently.

Concluding Remarks

Concluding remarks were offered by Sarah Maughan at the end of the seminar.

Sarah started by stressing the importance of a clear definition of purpose for any future national monitoring survey. This was a recurring theme throughout the seminar, and clarity of purpose is the essential foundation upon which any effective survey must be built.

Other decisions, for example about curriculum coverage and whether to subdivide outcomes by groups of pupils, will be determined by the purpose of the study.

Correspondingly, the decisions made about purpose, coverage and outcomes will give the information that will determine the size and composition of the sample.

Sarah reviewed the discussions of linkage between national and international surveys and found some encouraging signs that this proposal would be feasible in practice.

A final positive outcome from the seminar was the possibility that emerged of an integrated approach to the analysis challenges, rather than polarised debates pitching one methodology against another.

This seminar was set up in an attempt to move forward the technical debate about the real challenges in sampling and analysis that will face the proposed national monitoring survey. It is hoped that the outcomes from the discussion can be used to inform the future decisions that will be needed to establish a sound and robust methodology.

References

Beaton, A.E. and Gonzalez, E.J. (1997). 'Reporting achievement in mathematics and science content areas.' In: Martin, M.O. and Kelly, D.L. (Eds) *TIMSS Technical Report Volume II: Implementation and Analysis (Primary and Middle School Years)*. Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy [online]. Available: <http://timss.bc.edu/timss1995i/TIMSSPDF/TR2book.pdf> [19 April, 2010].

National Foundation for Educational Research (2009). *Submission to Expert Group: Issues to Consider when Developing a National Monitoring System (Version 2)*. Slough: NFER [online]. Available: <http://www.nfer.ac.uk/nfer/publications/99901/99901.pdf> [19 April 2010].

Parkes, C. and Maughan, S. (Eds) (2009). *Proceedings of the Policy and Research Seminar on Methods for Ensuring Reliability of Teacher Assessments: Tuesday 2 June 2009 at the Royal Institute of British Architects*. Slough: NFER [online]. Available: <http://www.nfer.ac.uk/publications/99902/> [19 April, 2010].

Sainsbury, M. and Maughan, S. (Eds) (2009). *Proceedings of the Policy and Research Seminar on National Assessment Arrangements for Key Stage 3: Friday 9 January 2009 at The Nuffield Foundation* [online]. Available: <http://www.nfer.ac.uk/publications/99903/> [19 April, 2010].

Appendix 1: List of participants

Delegates	Organisation
Sandra Johnson	Assessment Europe
Harvey Goldstein	University of Bristol
John Bell	Cambridge Assessment
Malcolm Hayes	Edexcel
Rose Clesham	Edexcel
Chris Wheadon	AQA
Ian Stockford	AQA
Mal Cooke	Scottish Government
Barbara Donahue	QCDA
Catharine Parkes	QCDA
Benedict Coffin	DCSF
Adam Hatton	DCSF
Lorna Bertrand	DCSF
Sarah Maughan	NFER
Marian Sainsbury	NFER
Chris Whetton	NFER
Graham Ruddock	NFER
Tom Benton	NFER
Dougal Hutchison	NFER
Naomi Rowe	NFER
Simon Rutt	NFER
Ben Styles	NFER

Appendix 2: National Monitoring: the context

Sarah Maughan



National Monitoring: the context

Sarah Maughan
January 2010



1

Recent Events

- **Abolition of key stage 3**
 - Expert Group on Assessment
 - *'A national sample testing system should be introduced for pupils at the end of Year 9, in order to monitor national standards over time.'*
- **Science at key stage 2**
 - Government response letter to Expert Group
 - *'To ensure national accountability for standards, I will ask the scientific community to help develop a national sample test of science for key stage 2.'*



2



Current Situation

- **Key stage 2 science**
 - 2010 and 2011: existing test papers with sample of pupils
 - 2012: new model for assessment
- **Key stage 3: English, maths and science**
 - 2011: pilot
 - 2012: new model introduced



3



Longer History

- **Assessment of Performance Unit**
 - Subjects: maths, language, science, modern foreign languages, D & T
 - Model: innovative use of assessment approaches – practical, group tasks, speaking, paper and pencil
 - Reporting: underachievement, standards over time, topics



4

International Lessons

- **National Assessment of Educational Performance, USA**
- **Scottish Survey of Achievement**
- **National Education Monitoring Project, New Zealand**
- **International surveys: PIRLS, TIMSS, PISA**

- *NFER: Developing a National Monitoring System*
- *Newton (2008): Monitoring national attainment standards*



5

Key Questions

- **Purpose**
- **Testing model**
- **Reporting requirements**


- **Technical aspects: sampling and analysis**



6

Appendix 3: Statistical Methods used for the Analysis of National Monitoring Surveys - Sampling

Simon Rutt, Dougal Hutchison & Harvey Goldstein



Statistical Methods used for the Analysis of National Monitoring Surveys

Sampling

Simon Rutt, Dougal Hutchison &
Harvey Goldstein

Thursday 28th January 2010



1

Sampling

Opportunity to put trust/confidence back into an assessment system. Many will try and criticise. Has to stand up to investigation

What are we sampling for;

Single analysis of national performance or combined with analysis of sub groups. What might we want in the future; gender, SEN, ethnicities

Will we require Item/AF analysis. Will we assess all areas of the KS3 maths curriculum? (using & applying, numbers, calculating, algebra, space shape & measure and handling data)



2

Sampling

What stratifiers – School or pupil. Prior attainment - how will SLTs affect the selection of pupils given that the last primary SLT may have been at L4 in June of year 5.

At what level – whole year group, whole class, within class (stigmatising pupils if results fall). Pupil sample could use NPD to select. Selecting at class level, do we know who we are getting?

Size of sample = what is the assessment for, what level of precision is required, what budget is available, what level of burden on schools is right for the project



3

EXAMPLES

- **APU**
- **Mathematics**
- **(English) Language**
- **Foreign Languages**
- **Science**
- **Design and Technology**

- **NAEP**



Also International Studies

- **TIMSS (Trends in Mathematics and Science)**
- **PIRLS (Progress in International Reading Literacy Study)**
- **PISA (Programme for International Student Assessment)**

- **Also**
- **IALS (International Adult Literacy)**
- **SAL (Scottish Adult Literacy Study)**



Looks like testing

- *But isn't exactly*
- **Enormous number of items**
- **Different test for each subject in a class**
- **Doesn't provide comparable individual scores**
- **Doesn't have the aim of comparing individuals (Low stakes)**
- **Doesn't identify individuals**
- **Not necessary to produce individual scores**



WHAT DO YOU WANT?

- **THE NUMBER**
- **Compare The Numbers (over time)**
- **Compare different subgroups (e.g. Gender, ethnic, etc)**
- **Different themes within the curriculum**
- **Overt secondary analysis**



EXAMPLES OF SAMPLE SIZE

- **TIMSS England 2007 Y4 c 4,300**
- **PIRLS England c4,400**
- **APU Maths c10,000**



HOW BIG A SAMPLE?

- Just looking at n of students
- A simple random sample of 400 students
95 percent confidence interval
- Mean ± 10 percent of its standard deviation
($1.96 \sigma / \sqrt{400}$)
- Proportion ± 10 percentage points



BUT it's not that simple

- *Test design TMI*
- PISA Maths 85 items
- TIMSS Grade 8 Maths 194 items
- APU Maths 650 items



HOW DO WE MANAGE THIS?

- A design in which all the pupils take the items?
- NO
- Matrix sampling design
- A DESIGN IN WHICH ALL THE ITEMS ARE TAKEN
- *But not by the same pupils*



LINKING Maths Science

Book	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12	M13	M14	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13	S14
1	■				■		■																					
2		■																										
3			■																									
4				■																								
5					■																							
6	■																											
7						■																						
8							■																					
9								■																				
10									■																			
11										■																		
12	■																											



SIZE OF SAMPLE

- Two important factors
- Design effect (Ratio of
- Var(actual sample) TO
- Var(simple random sample of the same size)

- Intraclass correlation rho(measure of similarity within clusters)
- $Deff = 1 + rho(b - 1)$
- b is cluster size



DEFF

- Even with quite small rho, sizeable b gives quite large Deff.

- $Rho = .05$
- $b = 31$
- $Deff = 1 + .05*(31 - 1)$
- $= 2.5$



SIZE OF SAMPLE

- *We could take 1 pupil per school.*
- Statistically this could be quite efficient (disregarding costs) if all we were interested in was pupil results
- Not good for relations with schools
- *You want to be able to*
- separate pupil and school effects
- Look at within-school mechanisms?



SOME KIND OF CLUSTERED SAMPLE

- Either:
- 2-stage: -schools: pupils
- 3-stage:-schools: classes: pupils



STRATIFICATION

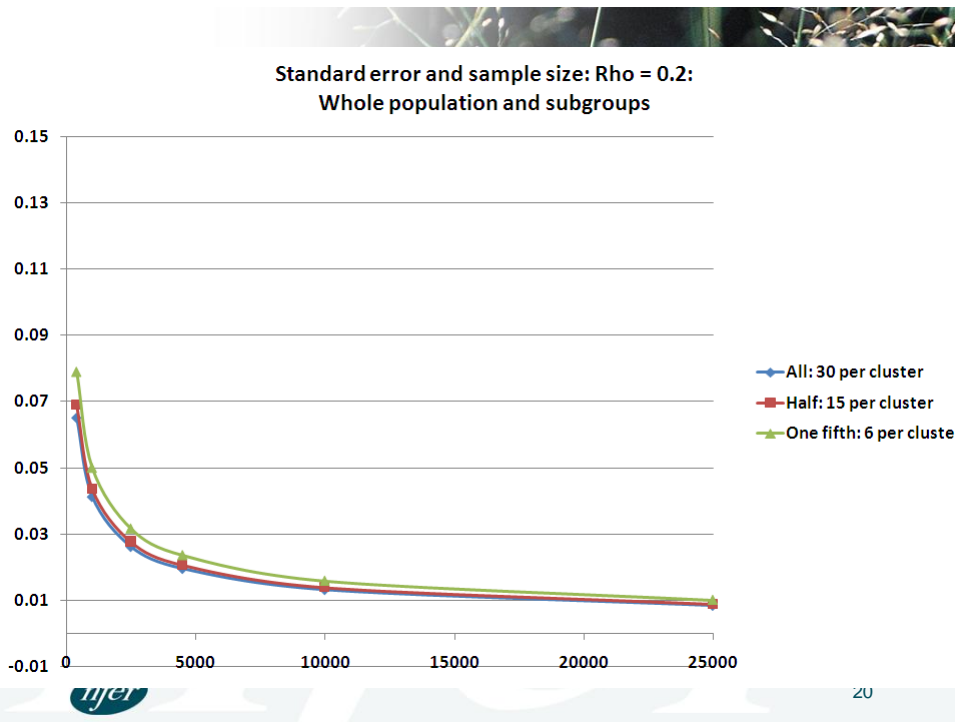
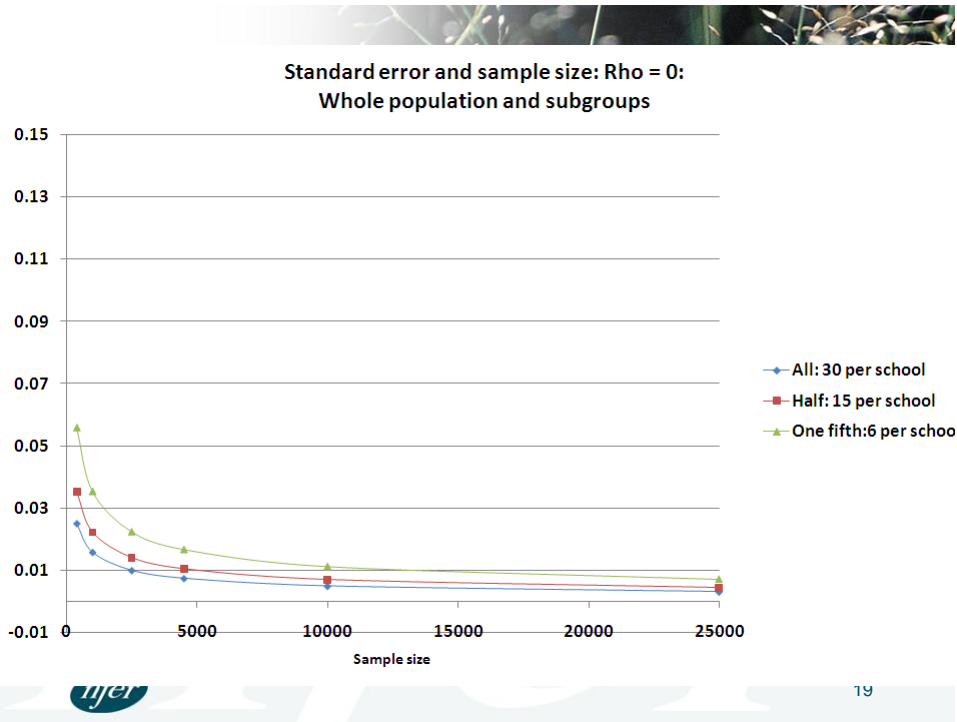
- **EXPLICIT**
- **Divide the population into separate groups**
- **E.g. State/Independent or Regions**
- **For which you could want separate estimates**
- **IMPLICIT**
- **Order population by some relevant variable e.g. National Curriculum results?**
- **And take a probability proportional to size (pps) sample**



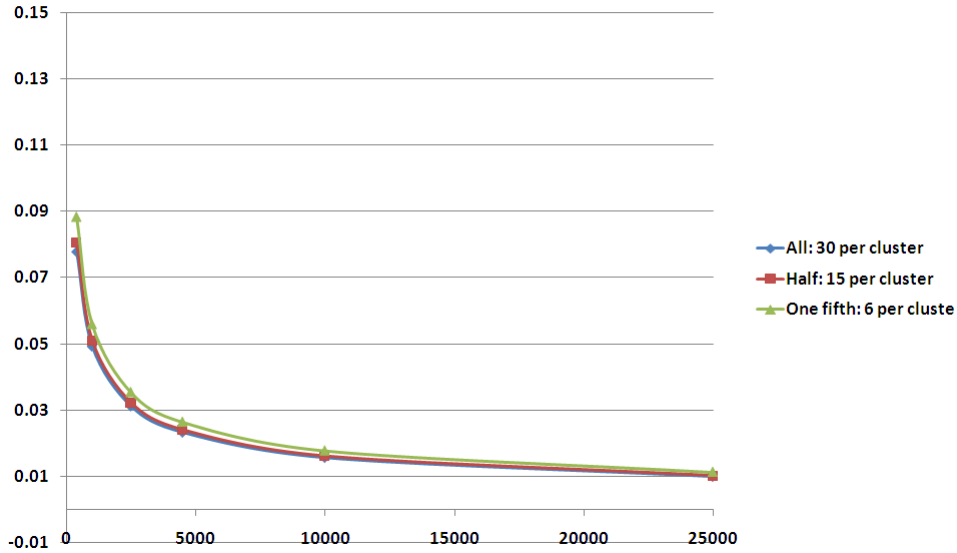
ALLOWING FOR NON-RESPONSE

- **1. Planned substitute schools**
- **2. Weighting of results to match population**
- **3. Modelling procedures.**



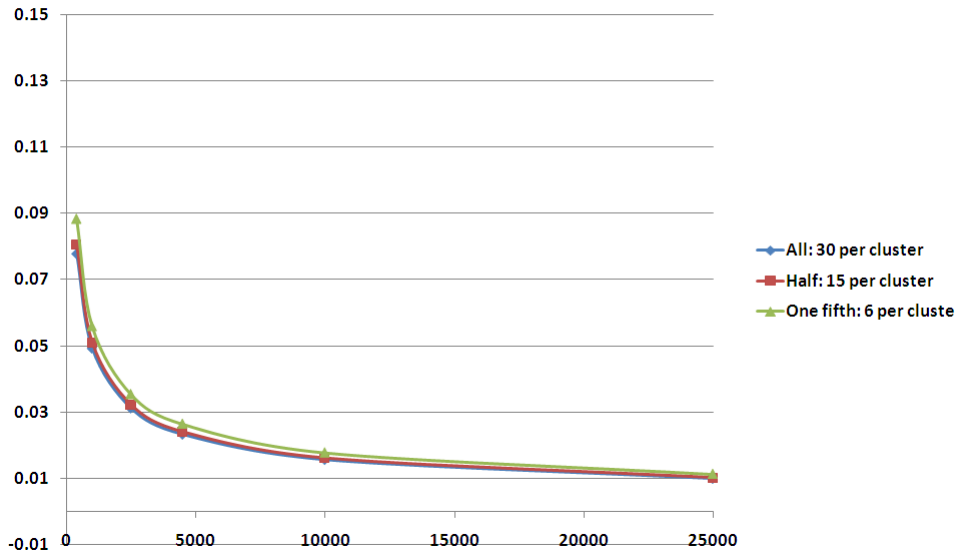


Standard error and sample size: $\rho = 0.3$:
Whole population and subgroups

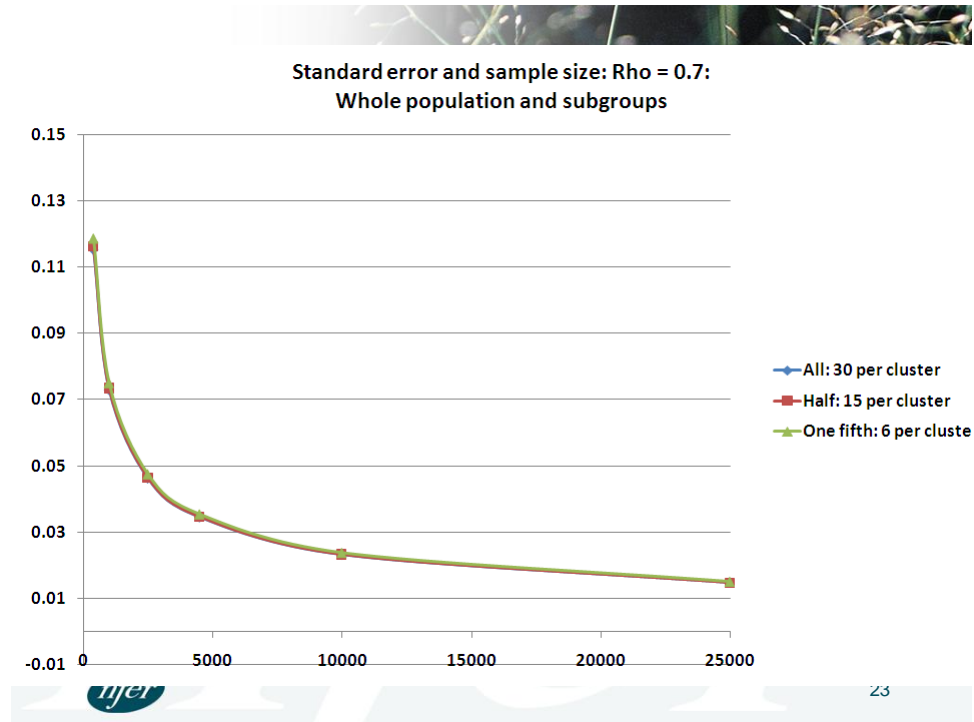


21

Standard error and sample size: $\rho = 0.3$:
Whole population and subgroups



22



- TO BE ABLE TO DESIGN THE SAMPLE....
- WE MUST KNOW WHAT WE WANT TO DO

nfer

nfer 24

Appendix 4: Recovering information in the NPD using multiple imputation

Harvey Goldstein & Tony Fielding

Recovering information in the
NPD using multiple imputation

Harvey Goldstein and Tony Fielding
University of Bristol

Sampling pupils at KS2

- Proposed that a sample only is tested at KS2
- Will not allow useful (value added) school effect estimates
- Can we still perform useful analyses?
- These need to use all the data efficiently

A missing data problem

- Propose the use of multiple imputation (MI) for the missing KS2 values.
- Note that all students have KS4 and this is fairly highly correlated with KS2
- This involves, for each missing value, sampling from its 'predicted' (posterior) distribution, conditional on (any) other variables
- We require that all the variables in the model of interest (MOI) enter imputation model and additional variables can also be used.
- Apply multiple imputation using REALCOM-IMPUTE (CMM website) that can handle 2-level and non-normal data.

The model of interest

- We use 1 year of NPD and choose KS4 (GCSE) score as outcome for all KS4 schools in East of England.
- Fit a 2-level model with covariates KS2, gender, FSM.
- Full data set results:

Full model

$$ks4_score_{ij} \sim N(\lambda B, \Omega)$$

$$ks4_score_{ij} = \beta_{0ij} \text{cons} + 0.189(0.005) \text{female}_{ij} + -0.237(0.010) \text{fsm}_{ij} + \beta_{3j} \text{ks2_score}_{ij}$$

$$\beta_{0ij} = -0.044(0.014) + u_{0ij} + e_{0ij}$$

$$\beta_{3j} = 0.682(0.005) + u_{3j}$$

$$\begin{bmatrix} u_{0ij} \\ u_{3j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.061(0.005) \\ 0.007(0.001) & 0.006(0.001) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.410(0.002) \end{bmatrix}$$

$$-2 * \log\text{likelihood(IGLS Deviance)} = 118512.875(60119 \text{ of } 60119 \text{ cases in use})$$

338 Schools, 60119 pupils. No allowance for mobility.

Sampling pupils

- Selecting a 3% sample which will yield ~15,000 pupils nationally
- Three analyses using MI:
 - Omit random 97% KS2 values giving 329 schools, 1746 pupils with KS2 values
 - Omit 83% KS2 values in 50 schools and 100% in remainder to give 1719 pupils with KS2 values
 - Omit 83% pupils in 50 schools and 100% in remainder to give 1719 pupils with KS2 values and add ethnic group and SEN status in imputation model

Table 1. Model of interest estimates: listwise deletion and multiple imputation. Standard errors in brackets. Response is KS4 score.

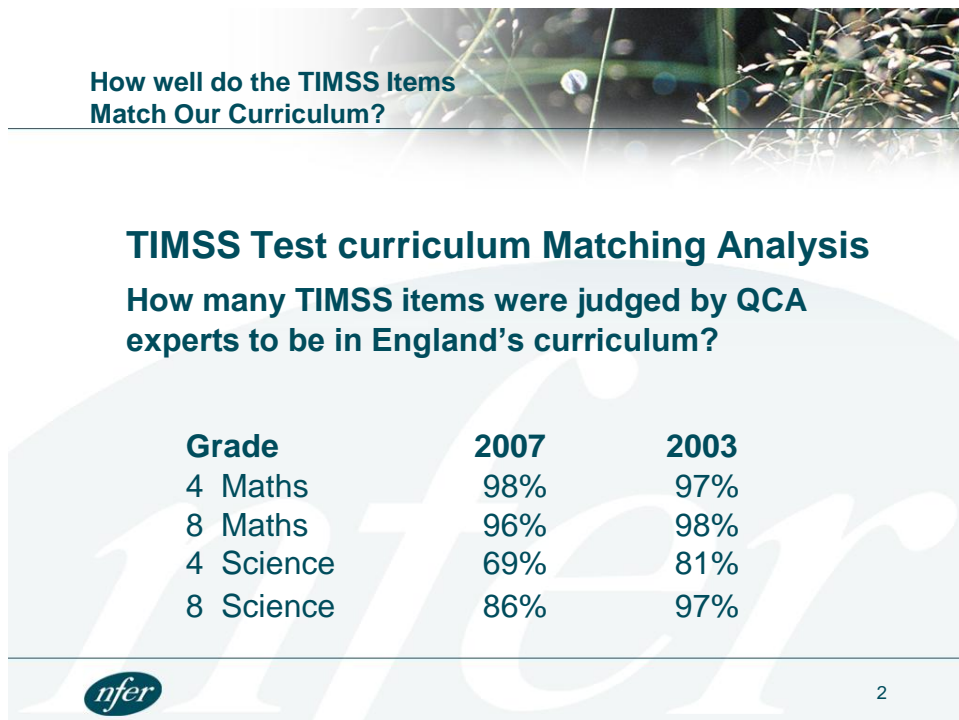
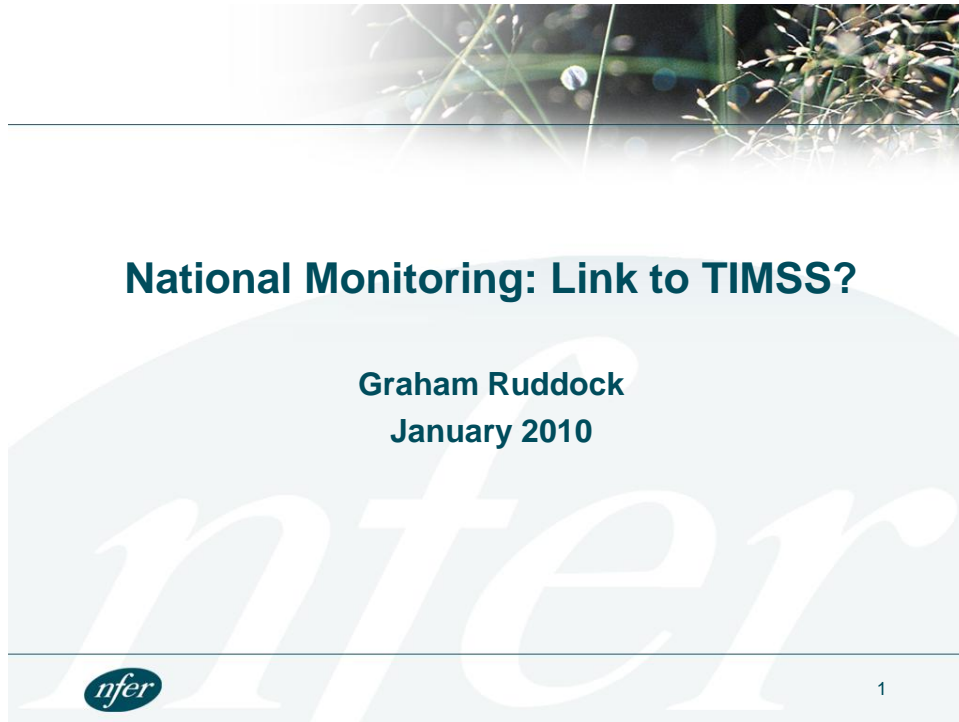
<i>Parameter</i>	Random 97% missing KS2		Random 83% missing in 50 schools, 100% elsewhere		Random 83% missing in 50 schools, 100% elsewhere + SEN, ethnic group imputation
	Listwise	MI	Listwise	MI	MI
<i>Intercept</i>	-0.079 (0.026)	-0.063 (0.021)	-0.037 (0.041)	-0.041 (0.031)	-0.048 (0.026)
<i>Female</i>	0.184(0.032)	0.200 (0.020)	0.172 (0.030)	0.188 (0.030)	0.192 (0.024)
<i>FSM</i>	-0.331 (0.061)	-0.272 (0.045)	-0.148 (0.054)	-0.130 (0.033)	-0.168 (0.051)
<i>KS2 score</i>	0.699 (0.017)	0.662 (0.014)	0.724 (0.022)	0.726 (0.015)	0.718 (0.012)
σ_{u0}^2	0.053 (0.011)	0.058 (0.005)	0.058 (0.014)	0.071 (0.009)	0.064 (0.007)
σ_{u01}	0.0 (0.0)	0.000 (0.001)	0.012 (0.006)	-0.001 (0.001)	-0.001 (0.001)
σ_{u1}^2	0.0 (0.0)	0.004 (0.001)	0.010 (0.005)	0.003 (0.001)	0.004 (0.001)
σ_{e0}^2	0.414 (0.015)	0.423 (0.013)	0.357 (0.013)	0.366 (0.011)	0.367 (0.011)

Implications

- For school effects a 3% sample yields 5 pupils per school on average with shrinkage of about 50% and CI factor of about 4.
- Sampling + efficient modelling using MI can provide useful analyses. More work needed on sample design and other auxiliary variables, including compositional ones, at pupil and school level.
- Further possibilities are sampling other variables possibly on rotating basis.
- Sampling most children in a school allows study of compositional effects.
- Took 11 hours on a Xeon 2.6 Ghz processor and 1 set of imputations for full data set with 3097 schools took 4 days.

Appendix 5: National Monitoring: Link to TIMSS?

Graham Ruddock



How well does the TIMSS Framework map to the English National Curriculum?

Science Work Carried out at NFER

Grade 4

Content domain	Assessed (%)	Not assessed (%)
Life science	92	8
Physical science	100	0
Earth science	40	60

Grade 8

Content domain	Assessed (%)	Not assessed (%)
Biology	75	25
Chemistry	83	17
Physics	95	5
Earth science	53	47



3

How well does the English National Curriculum map to the TIMSS framework?

Grade 4

English NC	Assessed (%)	Not assessed (%)
Sc1	73	27
Sc2	100	0
Sc3	100	0
Sc4	68	32

Grade 8

English NC	Assessed (%)	Not assessed (%)
Sc1	68	32
Sc2	76	24
Sc3	69	31
Sc4	75	25



4

Released TIMSS Items Only?

About half the TIMSS items are released after each survey. If these are regarded as not suitable for monitoring then only secure items should be used.

Items are in blocks of 12 or so, release is by block.



5

Structures

- **Year 9: Could link to the TIMSS IRT scale**
- **Year 9: Could link to the TIMSS benchmarks**
- **Year 6: Could link to TIMSS Grade 4, which is our Year 5**
- **TIMSS + local items?**



6

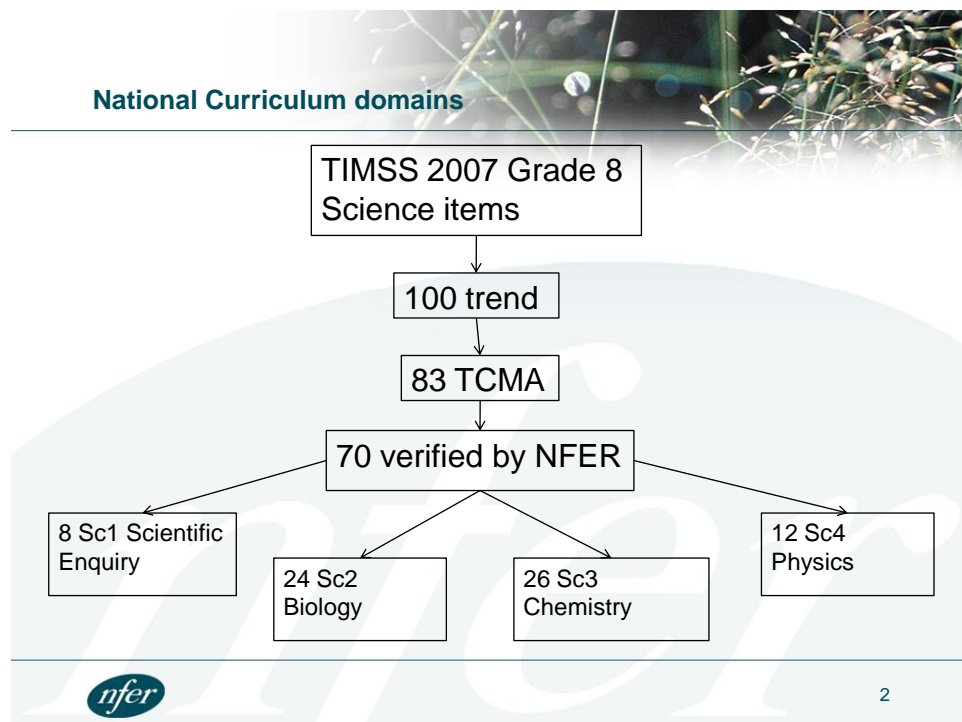
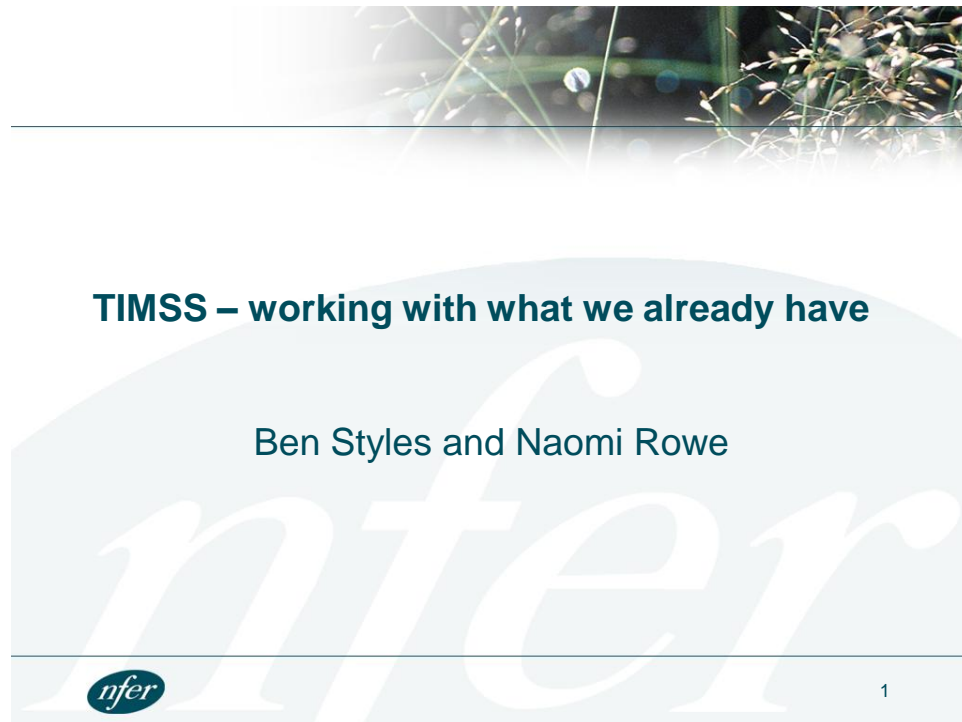
Structures

- **NC tests longer than TIMSS, so scope for “TIMSS +”**
- **With enough TIMSS items included to link and scale there are many possibilities**



Appendix 6: TIMSS – working with what we already have

Ben Styles & Naomi Rowe



Trend in average percent correct

Excerpt from TIMSS 2003 International Report

Countries	Average Percent Correct for Science Content Areas					
	Total Science Trend Items (74 Items)		Life Science Trend Items (17 Items)		Chemistry Trend Items (14 Items)	
	2003	1999	2003	1999	2003	1999
Australia	57 (0.7)	--	61 (0.8)	--	53 (0.9)	--
Belgium (Flemish)	56 (0.5)	60 (0.5) Ⓢ	61 (0.6)	64 (0.5) Ⓢ	49 (0.5)	51 (1.0) Ⓢ
Bulgaria	50 (1.1)	57 (1.1) Ⓢ	50 (1.2)	58 (1.3) Ⓢ	53 (1.2)	62 (1.1) Ⓢ
Chile	40 (0.5)	38 (0.7) Ⓢ	43 (0.6)	41 (0.8) Ⓢ	41 (0.7)	38 (0.7) Ⓢ
Chinese Taipei	62 (0.7)	67 (0.6)	63 (0.6)	68 (0.6)	74 (0.6)	73 (0.6)

Countries	Average Percent Correct for Science Content Areas					
	Physics Trend Items (22 Items)		Earth Science Trend Items (12 Items)		Environmental Science Trend Items (9 Items)	
	2003	1999	2003	1999	2003	1999
Australia	59 (0.9)	--	57 (1.0)	--	56 (1.0)	--
Belgium (Flemish)	61 (0.6)	64 (0.8) Ⓢ	56 (0.7)	59 (1.0) Ⓢ	49 (0.8)	54 (0.7) Ⓢ
Bulgaria	48 (1.1)	52 (1.4) Ⓢ	57 (1.3)	63 (1.2) Ⓢ	43 (1.3)	50 (1.3) Ⓢ
Chile	40 (0.5)	37 (0.7) Ⓢ	41 (0.6)	38 (0.7) Ⓢ	33 (0.6)	37 (0.8) Ⓢ
Chinese Taipei	62 (0.8)	64 (0.7)	69 (0.8)	71 (0.7)	70 (0.9)	69 (0.8)

Ⓢ 2003 significantly higher than 1999
 Ⓣ 2003 significantly lower than 1999



Methodology

- IEA methodology (TIMSS 1995 Technical Report Chapter 9 – Beaton and Gonzalez)
- Takes into account sample representativeness and sample design





Methodology - steps

- **Transform graded response items to a series of binary items**
- **Mean % correct across selected items calculated using the overall sampling weight**
- **Each replicate weight used to calculate a separate mean % correct**
- **Variation between original sample estimate and each replicate estimate is the sampling error**



5



The future

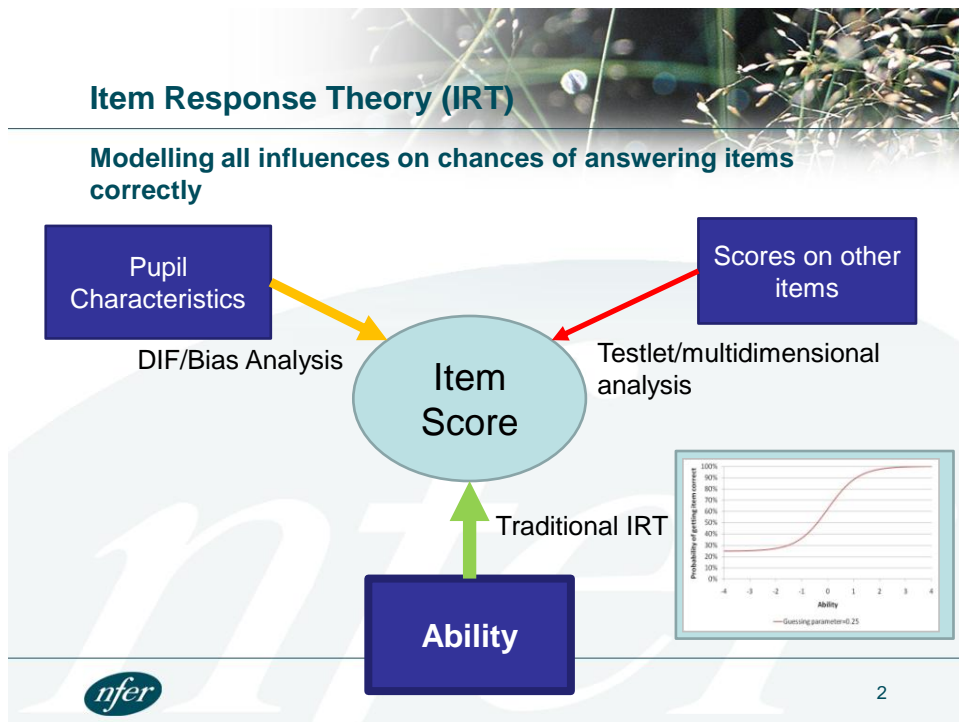
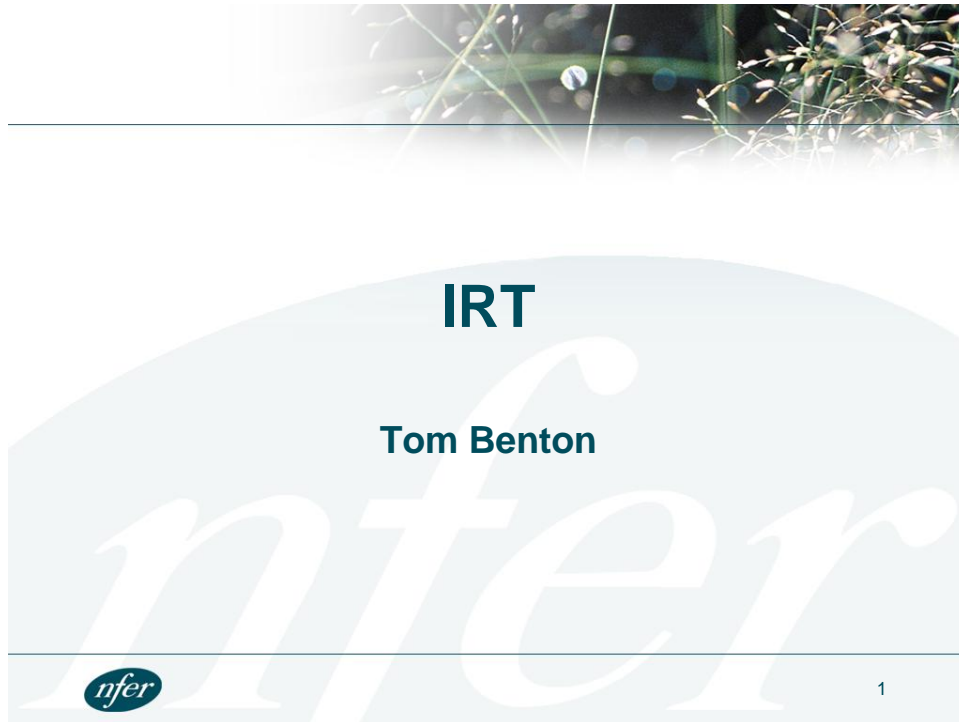
- **NFER internal project will show how trends in National Curriculum domains can be monitored using TIMSS**
- **Submitted to IEA International Research Conference 2010**
- **Technique could be used in both mathematics and science KS3 and KS2**



6

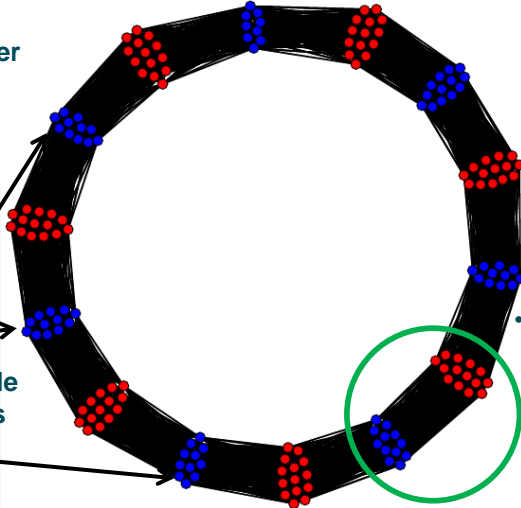
Appendix 7: Item Response Theory

Tom Benton



Advantages of IRT

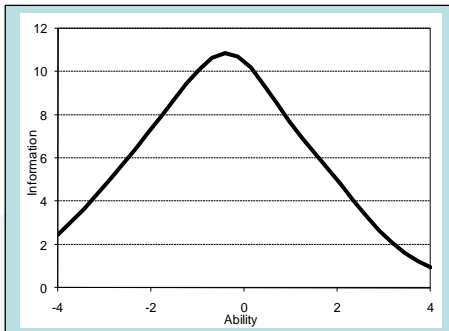
- Can use large number of items
- Good curriculum coverage
- A sample of items provide a link across years
- Each student only needs to attempt a few items



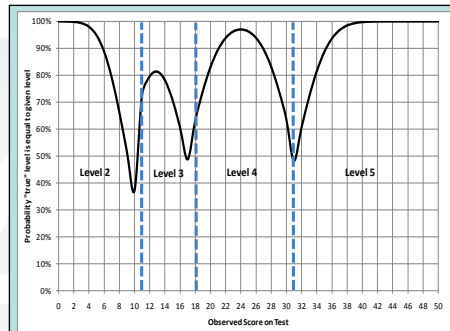
3

Applications of IRT Test Design

- Once we have item characteristics we can design a test with appropriate characteristics.



Continuous outcomes



Categorical outcomes



4

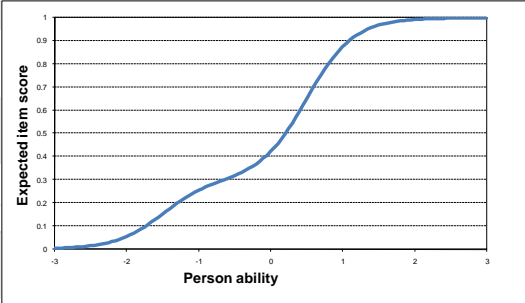
Applications of IRT Reporting results

- **National results**
 - Overall ability compared to previous years
 - Percentage at given thresholds
 - Standard errors for all estimates taking account of:
 - Sampling Error
 - Measurement Error
- **Student Results**
 - Student results on a common scale (with confidence intervals)



5

Alternative IRT models and assumptions

Model	Basic idea
1 Parameter	Every item has equally strong relationship with
2 Parameter	
3 Parameter	
4 Parameter	
Semi-parametric	
Multi-dimensional	...and some groups of items might be inter-related...
Testlet Response Theory	...and some other individual items might be related to one another

Where do I draw the line?



6

Sensitivity analysis

- **Results should be determined by:**
 - performance of students
 - not by favourite model of statisticians
- **Sensible limits on required accuracy**
 - How much difference is there between different methods



7

Sensitivity analysis – Different models

- **Compare different models for TIMSS grade 4 maths**
 - Basic 2 parameter IRT
 - Testlet response model
 - (applied to 5 pairs of items showing greatest residual covariance)
 - Semi-parametric model
 - (applied to 5 items with greatest lack of fit)



8

Sensitivity analysis – Different models

Model	Estimated percentage of students above 2003 median achievement	Difference between 2003 and 2007 (scale mean 500, SD=100)
2 Parameter IRT	49.18%	-1.07
Testlet Response Model	49.41%	-0.70
Semi-parametric model	49.20%	-1.02



9

Summary

- **Powerful and growing methodology**
- **Must be aware of assumptions**
- **Quoted accuracy should not exceed sensitivity of method**
 - If test fits model assumptions then this probably won't matter
 - Effort on test design avoids problems later



10

Appendix 8: Domain sampling and generalizability theory

Sandra Johnson

Domain sampling and generalizability theory

Sandra Johnson, Assessment Europe
NER Seminar, January 2010

Aim(s) of the future Key Stage 3 sample-based survey programme

- to estimate population and population subgroup attainment in key subject areas, and to monitor the situation over time (principal aim?)?
- to gather contextual information about teaching and learning, including provision, and to monitor this over time?
- to provide pupil-level attainment estimates for some purpose?

The pupil population

- all pupils nearing the end of Key Stage 3 (with some defined subgroup exceptions)
- the pupils will be “nested” in classes and schools, and in gender and deprivation group
- pupil sampling will be multi-stage (complex)

What to assess as population attainment or achievement, and how to assess and monitor it?

Sandra Johnson, NFER Seminar 2010

The subject domain

- essentially the Key Stage 3 curriculum, comprising some combination of subject knowledge, understanding and skill
- must be operationalised in terms of “valid” test questions
- to guide operationalisation question descriptors are useful

Unlike the pupil population, the subject domain does not typically pre-exist as a “population” of physically countable elements.

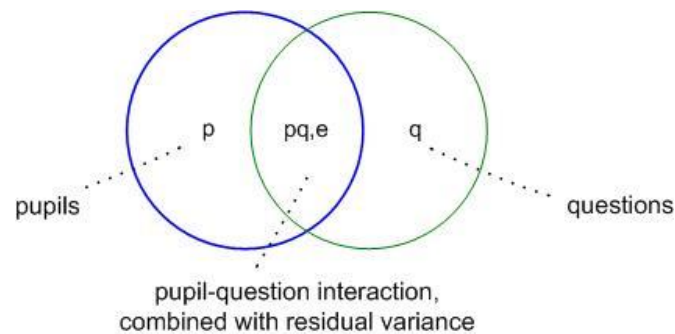
Sandra Johnson, NFER Seminar 2010

Sampling and estimation:

- sample the pupil population (typically complex sampling)
- sample the subject domain, i.e. the question pool
- use matrix sampling to distribute the sampled test questions among the sampled pupils
- produce an appropriate domain-referenced performance measure for the set of survey pupils, and estimate its precision as a population estimate

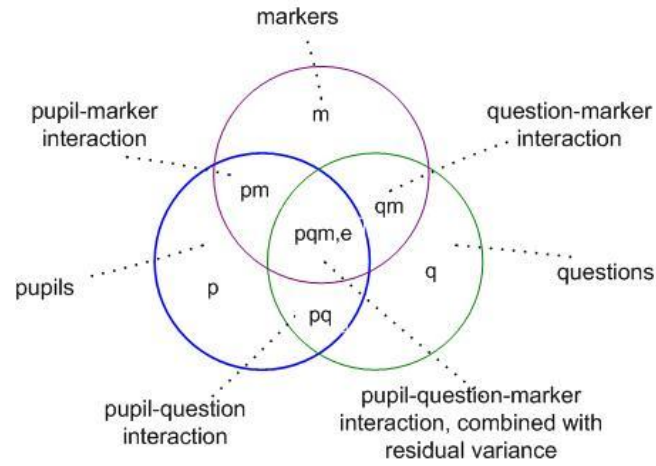
Sandra Johnson, NFER Seminar 2010

When pupils meet questions...



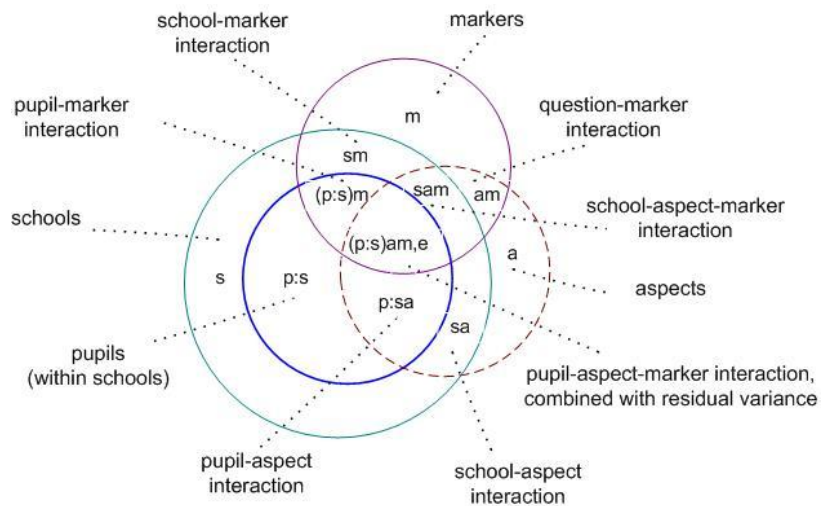
Sandra Johnson, NFER Seminar 2010

and human marking is unavoidable...



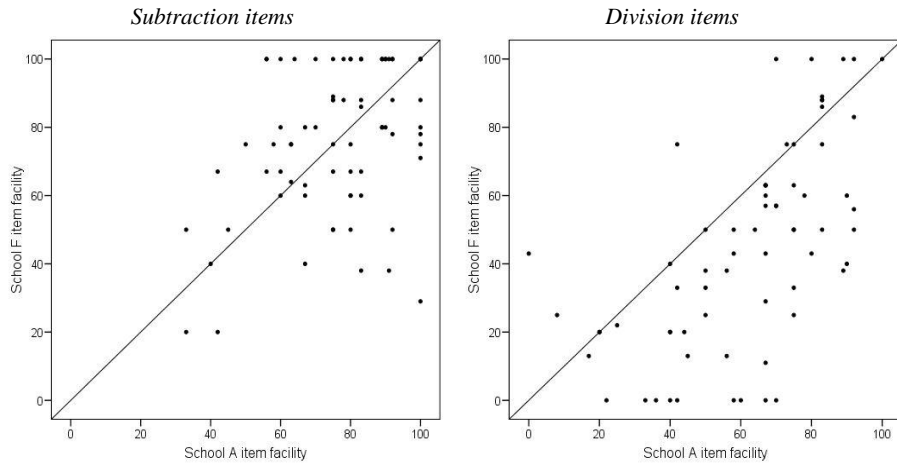
Sandra Johnson, NFER Seminar 2010

Recognising hierarchies...



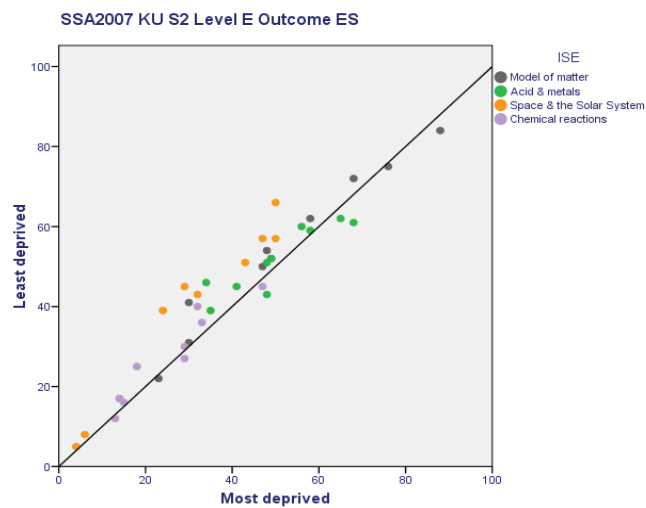
Sandra Johnson, NFER Seminar 2010

6-school research study (2008/09) (LHS typical picture, RHS division weakness for school F)



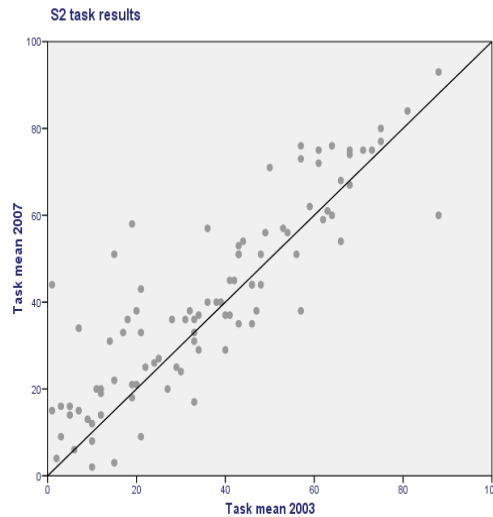
Sandra Johnson, NFER Seminar 2010

Subgroup differences in mean task scores (DIF) (Scotland, science at S2)



Sandra Johnson, NFER Seminar 2010

Changes in mean task scores over time (Scotland, science at S2)



Sandra Johnson, NFER

Principal strengths of the domain sampling approach

- validity
- simplicity
- transparency
- comprehensibility

Sandra Johnson, NFER Seminar 2010