February 2013

Good Assessment by Design

An International Comparative Analysis of Science and Mathematics Assessments

Dr Rose Clesham

ALWAYS LEARNING

PEARSON

International Comparative analysis of Mathematics and Science assessments

Introduction

The desire to compare national standards (nationally over time, and against high-performing countries) in the core curriculum areas of Mathematics and Science has led to increasing critical analysis of the structure and function of the educational system in the UK, and their associated curricula and qualifications. National primary and lower secondary programmes of study have been revised over the years, and now the main academic qualification pathway at KS4 (GCSE) is being reviewed, with potentially significant amendments made to both content and performance standards.

The purpose of this international comparative assessment research was to compare the assessment instruments of selected countries and jurisdictions with those of England in the core subject areas of Mathematics and Science (including Biology, Chemistry and Physics and Science) for 16-year-olds, and required the development and use of a systematic and empirical methodology in order to objectively evaluate the comparative breadth and depth of different jurisdictions' summative examinations.

The objectives of this research can be summarised as follows:

- to provide comparisons between UK-based assessments and selected high-performing PISA jurisdictions
- to analyse assessment instruments to investigate how content standards were operationalised and the type and demand of cognitive operations employed
- to provide robust comparative indicators on the nature and demand of a number of national and international assessments for students at approximately 16 years of age (or nearest equivalent)
- to provide examples of good assessment practice where such evidence is found
- to inform the structures and demand of new GCSE assessment structures to ensure high quality assessment outcomes by design.

The countries/jurisdictions were selected according to the following criteria:

- They perform highly on international benchmarking assessments, e.g. TIMSS and PISA.
- They have shown significant improvement in benchmarks over the last 5 years.
- They have developed innovative good practice in their assessment systems.
- They have aspects of their assessment systems that are similar and therefore are useful comparators to the UK GCSE.
- In the main, their assessments are in the medium of English.
- The assessments are summative and linear.
- UK qualifications at the same Key Stage were included.

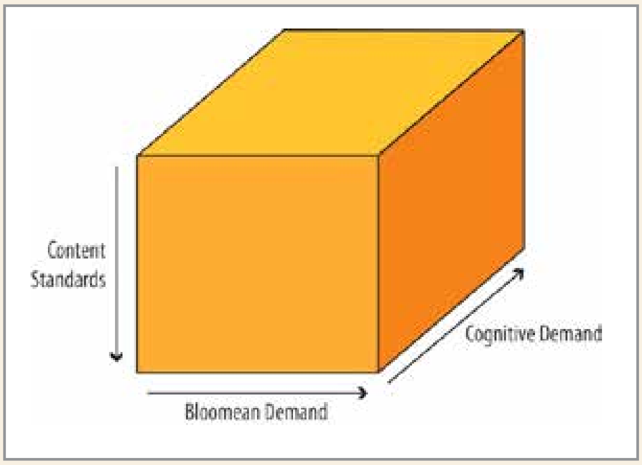
For the purposes of this paper, the examination boards and jurisdictions have been anonymised. They included externally marked summative assessments from UK examination boards (national and international versions) and assessments from Australia, Canada, US, Singapore, Hong Kong and PISA. Without the associated names, it may be difficult to make active comparisons, however areas A–E were UK-designed assessments.

In terms of comparing, analysing and evaluating the assessment instruments themselves, the following aspects were investigated:

1. Content mapping against the content standards.
2. The type of cognitive operation for each question item (based on Bloom (1956) and Porter (2002)).
3. The cognitive demand of each item.
4. The item types (eg. MCQ, short numeric, short-open, long-open or extended response).
5. The number of marks awarded for each of the above categories of item type.

The first three aspects form a three-dimensional mapping instrument that provide a framework for profiling how content standards are sampled, represented and assessed. The last two aspects provide additional useful information in the evaluation and design of assessments. The three-dimensional cube model (Clesham, 2012a, 2012b) is shown and described below. This model has been used in the analysis of a number of national and international assessments and applied to the new PISA scientific literacy framework for 2015 (OECD, 2012).

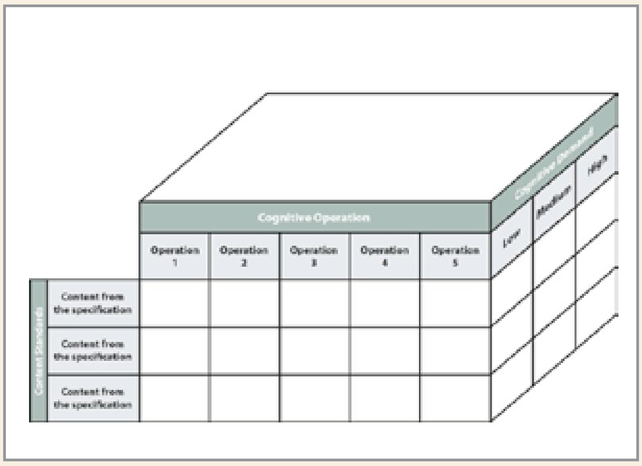
Figure 1: Three-dimensional cube model



The cube provides a framework for mapping items against content standards and cognitive operations (these are effectively a proxy for Bloom) and also mapping using a third dimension based on a cognitive demand taxonomy. This provides a means of evaluating and designing assessments using a number of key criteria.

This conceptual framework was operationalised by systematically mapping assessments item by item, on each of the three axes, which then allowed for a series of analyses to be performed.

Figure 2: Mapping assessments



The three axes are now described.

The Content Standards

This axis describes the intended focus of a course of study in terms of the subject content, any identified skills or competencies and their relative emphases. These are described in a number of ways internationally and therefore there are direct comparability issues. Countries operating on a K–12 basis usually develop yearly content standards, while other countries (England included) develop educational phase content standards. KS4/GCSE, for example, covers two years of study. The purpose and style of content standards also vary internationally, often linked to the manner in which they are assessed. Some countries use content standards to be definitive statements of teaching requirements and prescribe the topics and subtopic areas to be taught and to what Bloomean level of demand (**state**, **describe**, **explain** etc.). Other countries write content standards to be more indicative, therefore with less prescribed detail. Sometimes verbs are not used at all and sometimes generic interpretations of Bloomean verbs are used (e.g. ***explain*** subsumes any questions that ask for lower Bloomean hierachies, e.g. recall and descriptions). An intention of less prescribed content standards is to avoid a narrowly defined curriculum, and associated 'teaching to the test'.

For the purposes of making international comparisons, a set of uniform content standards was developed for each subject, as described by Porter (2002) and used in the DfE research study (2012). These provided a framework for mapping national or international assessments in terms of the way the assessments covered the curriculum content. The detail of the way content standards were written or described was also used as one consideration of cognitive demand. This issue is dealt with in the section describing the approach to mapping cognitive demand.

The Bloomean demand

This axis describes the manner in which assessment tasks or items are presented to students and the intended nature of their responses. This axis of the cube can be adapted depending on the most useful taxonomy for the assessment systems concerned (e.g. assessment objectives or competencies), however as this research was focused on the analysis of a number of disparate international systems, the use of cognitive operations was considered to be the most appropriate taxonomy. Porter (2002) used these categories to describe cognitive demand. This research supposes that these categories are best used to describe the nature of test tasks or items, but that they do not necessarily imply absolute categories of demand. This issue will be explored further in the analyses of the data.

The cognitive operations with their associated descriptions, used for mapping Mathematics and the sciences were:

Mathematics

Memorise

- Recite basic mathematical facts; recall mathematical terms and definitions; recall formulae and computational processes.

Perform procedures

- Use numbers to count, order or denote; do computational procedures or algorithms; follow procedures/instructions; make measurements, do computations; solve equations/formulae, routine word problems; organise or display data; read or produce graphs and tables; execute geometric constructions.

Demonstrate understanding

- Communicate new mathematical ideas; use representations to model mathematical ideas; explain findings and results from data analysis; develop/explain relationships between concepts; explain relationships between models, diagrams and other representations.

Conjecture, generalise, prove

- Determine the truth of a mathematical pattern or proposition; write formal or informal proofs; analyse data; find a mathematical rule to generate a pattern or number sequence; reason inductively or deductively; use spatial reasoning.

Solve non-routine problems, make connections

- Apply and adapt a range of appropriate strategies to solve problems; apply mathematics outside of mathematics contexts; recognise, generate or create patterns; synthesise content and ideas from several sources.

Science

Memorise

- Recite basic science facts; terms and definitions; recall scientific formulae.

Perform procedures

- Make observations; collect and record data; use appropriate tools; make measurements; do computations; organise and display data in tables or charts; execute procedures; conduct experiments; generate questions, make predictions; test effects of different variables.

Communicate understanding

- Explain concepts; observe and explain teacher/student demonstrations; explain procedures and methods of science and enquiry; organise and display data in tables or charts; student presentations of scienctific information.

Analyse information

- Classify and compare data; analyse data, recognise patterns; reason inductively or deductively; draw conclusions; identify faulty arguments or misrepresentations of data.

Apply concepts/make connections

- Apply and adapt science information to real-world situations; apply science ideas outside the context of science; build or revise theory/plan and design experiments; synthesise content and ideas from several sources; use and integrate scientific concepts.

Cognitive demand

This axis describes the type of knowledge or processing required in assessment tasks or items. It is the category that required the most careful consideration in its conceptualisation, and has a significant impact on the structure, design and evaluation of content standards, assessment items and associated marking/scoring rubrics.

In most assessment frameworks, item or test difficulty, usually empirically derived, is often confused with cognitive demand. Difficulty usually refers to the amount of knowledge to answer an item, whereas cognitive demand refers to the type of processing required (Davis and Buckendahl, 2011). Thus items that merely require recall of one piece of information make low cognitive demands, even though the knowledge itself might be quite complex. In contrast, items that require recall of more than one piece of knowledge and require a comparison and evaluation of the competing merits of their relevance would be seen as having high cognitive demand. The cognitive demand of any item, therefore, includes the degree of complexity and range of knowledge it requires, the cognitive operations that are required to process it and the level of specificity related to the manner which content standards are described and assessed.

Care therefore needs to be taken to ensure that the relationship between the item difficulty, cognitive operations and cognitive demand are understood more explicitly by test developers and users of national and international assessments. Tests need to provide a range of cognitive operations and demands for students across the ability range (Brookhart and Nitko, 2011), as well as items based on empirical data alone (item difficulties).

Background to the cognitive demand methodology

The methodology used in this study to evaluate the cognitive demand of the assessments is based on a number of schema, which have all been used to consider cognitive demand in assessment instruments. Before describing the criteria for demand that were used, a brief description and discussion of contributing schema are given.

Various classifications of cognitive demand schemes have been developed and evaluated since Bloom's Taxonomy was first published (Bloom, 1956). These have been largely based on categorisations of knowledge types and associated cognitive processes that are used to describe educational objectives or assessment tasks. These include the works of Anderson and Krathwohl, 2001; Marzano and Kendall, 2007; Ford and Wargo, 2011.

Another schema can be found in the framework based on "Depth of Knowledge" developed by Webb (1997) specifically to address the disparity between assessments and the expectations of student learning. For Webb, levels of depth can be determined by taking into account the complexity of both the content and the task required. His schema consists of four major categories: level 1 (recall), level 2 (using skills and/or conceptual knowledge), level 3 (strategic thinking), and level 4 (extended thinking). Each category is populated with a large number of verbs that can be used to describe cognitive processes. Some of these appear at more than one level. This framework offers a more holistic view of learning and assessment tasks and requires an analysis of both the content and cognitive process demanded by any task. Webb's depth of knowledge (DOK) approach is a simpler but more operational version of the SOLO Taxonomy (Biggs and Collis, 1982) which describes a continuum of student understanding through five distinct stages of pre-structural, uni-structural, multi-structural, relational and extended abstract understanding.

Finally, the CRAS framework (Pollitt, 2007) was considered and forms a constituent element of some of the conditions used to determine cognitive demand. The CRAS framework was created as an adaptation of an earlier scale of cognitive demand developed by Edwards and Dall'Alba (1981) to evaluate lower secondary science materials. It uses aggregates of complexity (C), resources (R), abstractness (A) and strategy (S) to produce a CRAS rating of item and test demand. However, as assessments often do not utilise all four CRAS strands, and the overall rating is based on the averages of item averages, the scale itself is limited in its application to operational assessment design or evaluation principles.

All of the above frameworks deal in different ways with the issue of cognitive demand in assessment instruments. The third dimension of the cube framework shown above has therefore been informed by various elements of Bloom (1956), Webb (1997) and Pollitt (2007) to produce a useable tool to objectively map cognitive demand, which is a constituent part of the evaluation or design of assessments.

The factors that indicated the cognitive demand of items and/or mark schemes included:

- The complexity of elements of knowledge or task–linked to the expectation of the content standards of the qualification level.
- The number of steps/linkages involved in a response.
- The level of familiarity/prior knowledge students may have of the content or procedures required. This also relates to the expectations set out in the specification, or whether the procedure is routine and does not require any adaptation or application. For example if the specification contains all the detail required to answer a question, this will affect the demand rating.
- The predictability of a question-from series to series–how familiar the question is over time.
- The manner in which marks are awarded – **this is important**. Even if a question is demanding in terms of what a student is expected to do, if the mark scheme breaks down the marks very atomistically so that the student does not need to carry out linked steps. This will affect a demand rating.
- The use of verbs or command words – this is clearly a factor – but very dependent on the previous five above.

Therefore in order to make their judgement on H/M/L, the raters considered not only the question paper, but also the way the mark scheme and content standards document were constructed and described. Although it was an important element for raters to discuss, agree and standardise on contributing criteria, and the descriptors expressed some subject-specific focus or language, there was a consistent cross-subject underpinning idea of the rationale for L/M/H judgements.

The following statements are examples of how Low, Medium and High Cognitive demand were described across subjects.

Low (L)

- Carrying out a one-step procedure, for example recall or description of a fact or facts that are not linked, a term, principle or concept.
- Recall or describe a fact or facts, a term, principle, definition, concept or procedure.
- A single step, routine question or simple, set procedure.
- Apply a formula (one-step).
- Apply a standard set of conventions or criteria.
- Locate information in a table or diagram.

Medium (M)

Requires some mental processing beyond a habitual response, that is, a decision has to be made by the candidate.

- Use and application of conceptual knowledge to describe or explain phenomena that required linkage, select appropriate procedures involving two or more linked steps, organise/display data, interpret or use data or graphs.
- Negotiate a route through a problem, that is, there are points where candidates can get stuck.
- Use and apply conceptual knowledge to describe or explain phenomena that require linkage.
- Select appropriate procedures involving two or more linked steps.
- Demonstrate the use of knowledge. Apply a set of rules or protocols to a situation.
- Identify cause and effect.
- Apply a skill or concept to a task.
- Interpret or use data or graphs. Convert information from one form to another.
- Multi-step calculations.

High (H)

Cognitive demands are more complex and abstract. Candidates need to consider a number of ideas and transfer knowledge.

- Analyse complex information or data. Use evidence. Draw conclusions from observations.
- Use concepts to solve problems. Work with complex concepts.
- Synthesise or evaluate evidence. Use conjecture. Make predictions. Justify.
- Analyse complex information or data, synthesise or evaluate evidence, justify, reason, giving various sources demonstrating complex linkage, develop a plan or sequence of steps to approach a problem
- Reason, giving various sources demonstrating complex linkage.
- Develop logical argument from a concept.
- Develop a complex plan or sequence of steps to approach a problem.
- Complete complex calculations with multiple processes.

Applying the three-dimensional mapping methodology

Subject expert raters were established for Mathematics, Science, Biology, Chemistry and Physics. These groups consisted of senior examiners, expert and experienced at writing and reviewing both content standards and assessment materials. These subject expert raters attended a day of guidance and training on the methodology in order to agree and standardise subject-specific categorisations and definitions to be used in the research.

All national and international content standards and assessments used in this research were provided to the raters in paper form, and all their analyses were collected in bespoke spreadsheets, which then allowed a number of analytical functions to be carried out. Space was also provided for the raters to make comments on any items that were difficult to categorise or had perceived structural or quality issues, good or bad.

All assessment comparisons were carried out by at least two raters, using a minimum of one exam series from each national or international assessment, and subject judgements were aggregated on the basis of group expert judgement rather than a 'true score'. Each of the expert subject rater groups agreed on the uniform content standard categorisations and the key indicators of cognitive operations and cognitive demand for their subject area. Although not a constituent part of content, cognitive operations or demand mapping, the raters also collected item information in terms of question types and the number of marks used for items.

Mapping results and analyses

This section of the report illustrates and discusses the analyses of the assessment-mapping exercises. As described earlier, there were five mapping categories collected in every assessment series from each examination board/jurisdiction:

1. Content mapping against the generic international content standards.
2. The type of cognitive operation for each question item (based on Bloom).
3. The cognitive demand of all assessment items.
4. The question type for each item (eg. MCQ, short numeric, short-open, long-open or extended response.
5. The number of marks for each item.

From these categories, ten sets of comparative analyses were carried out in Mathematics, Biology, Chemistry, Physics and Science. The first four analyses show how the assessment of content, cognitive operations, cognitive demand and question type compare across examination boards and jurisdictions. The next six analyses show correlations between these categories:

6. Cognitive operation v cognitive demand
7. Cognitive operation v question type
8. Question type v cognitive demand
9. Content coverage v cognitive operation
10. Content coverage v cognitive demand
11. Content coverage v question type.

The number of marks for each of these criteria were then used to carry out a comparative alignment exercise using Porter's Alignment Indices (Porter, 2002):

$$Alignment\_index = 1 - \frac{\sum |X-Y|}{2}$$

Where X = all the cell proportions in one matrix (e.g. cognitive operations)
And Y = all the cell proportions in the other matrices (cognitive demand).

The Porter alignment index is a simple measure of the distance between two matrices of proportions. So, for example, if two assessments are evaluated – the dimensions of cognitive operations and cognitive demand – and matrices completed to provide the proportions of marks allocated to each cross-classification, then the Porter alignment index can be computed as one minus half the sum of the absolute differences between corresponding cells. The index takes values between 0 (total misalignment) and 1 (exactly matching proportions in every cell).

In this research study, alignment indices have been calculated on the first three correlations (5, 6 and 7) across all jurisdictions and tabulated. It is important to note that these indices do not imply any good or bad value, only alignment, which may be judged beneficial or not.

The tables themselves are difficult to interpret, so each table has been subjected to a hierarchical cluster analysis, from which a dendrogram has been created. The method used was to first convert the matrix of alignments to a distance matrix, based not only on the alignment between each jurisdiction but also on the similarity of their alignment with every other jurisdiction. From the distance matrix, hierarchical cluster analysis using Euclidean distances as the proximity measure was carried out and the results used to construct dendrograms.

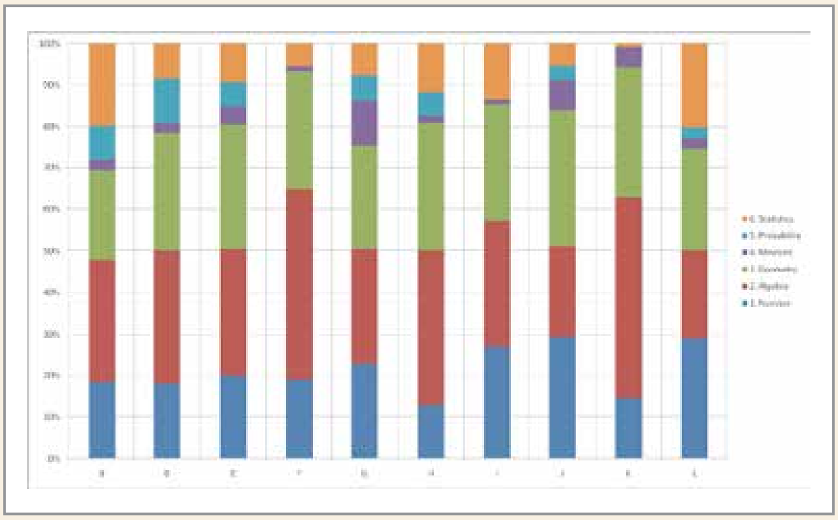
These dendrograms illustrate the relationship between the jurisdictions in terms of proximities and can be interpreted as the nearer to the bottom of the diagram jurisdictions are joined, the more similar they are. The height of a clustering gives an indication of how far apart the jurisdictions or cluster of jurisdictions are; horizontal distances do not convey meaning.

As content coverage is usually sampled over assessment series, particularly in terms of sub-categories of content standards, it may not be valid to discuss in too much detail correlations 8, 9 and 10 across exam boards and jurisdictions. In this section of the report, the first seven analyses will be shown and discussed in Mathematics and Biology, Chemistry, Physics and Science. The other remaining analyses from all subject areas are shown in the appendices (not included in this report).

For the purpose of this report all the examination boards and jurisdictions are identified by the letters A–L. As there are more international jurisdictions that have summative assessments at the age of 16 in Mathematics and Science (including PISA assessments), these subject areas have more comparative judgements than the separate science subjects of Biology, Chemistry and Physics. Where there were separate tiered assessments, rather than amalgamating papers that were set for differing profiles of students, only the results from the higher-tier papers are shown in the following charts.

Mapping outcomes
Mathematics international comparisons

Table 1: Content standards



(A) and (C) were not used in the Mathematics comparative analysis as the only assessments available from them were modular. One of the criteria for this international comparative study was to use only linear examinations, to ensure that there were consistent and fair comparisons between differing systems. The content areas were established using generic uniform categories. Each of these subject areas contained a number of sub-topics which were also mapped.

It can be clearly seen from Table 1 that there was considerable variation on proportions of questions from the six Mathematics content areas across national and international assessments. It must be noted that usually linear assessments use sampling strategies over time to ensure content representation. As this study was purposed to look across a number of examination boards and jurisdictions over one or two examination series, the content coverage may not be generalisable. However, on the basis of this evidence there are significant differences in the emphases of different aspects of Mathematics assessed in examinations. If sampling strategies are employed, there are two potential models that may be used. One is to create a proportion of items in each mathematical content area and then maintain that proportion in every examination, the model is to vary the proportions across examination series in order to reduce predictability. How differing examination boards or jurisdictions create or maintain a sampling strategy with an associated justification and rationale is a key assessment issue in terms of validity, reliability, standard setting and maintenance. There is no assumption here that content areas should be equally weighted, and it is highly likely that some content areas are larger and take more curriculum time than others. This, however, should be verified.

It can be seen for Table 1 that in the main, the majority of examinations focused on number, algebra and geometry in large but in varying proportions. Number ranged from 13% (H) to 29% (L), algebra from 22% (J) to 46% (F) and geometry from 21% (B) to 32% (J). Measure, probability and statistics were generally minor assessment areas, and in some cases not assessed at all. Measure ranged from 1% (I) to 10% (G), probability from 0% (F, I and K) to 10% (D) and statistics from 1% (K) to 21% (L). G covered all content areas in the most even proportions. Although PISA do not strictly have established content standards, raters did not have difficulty in mapping PISA items using the uniform content standards.

Table 2: Cognitive operation

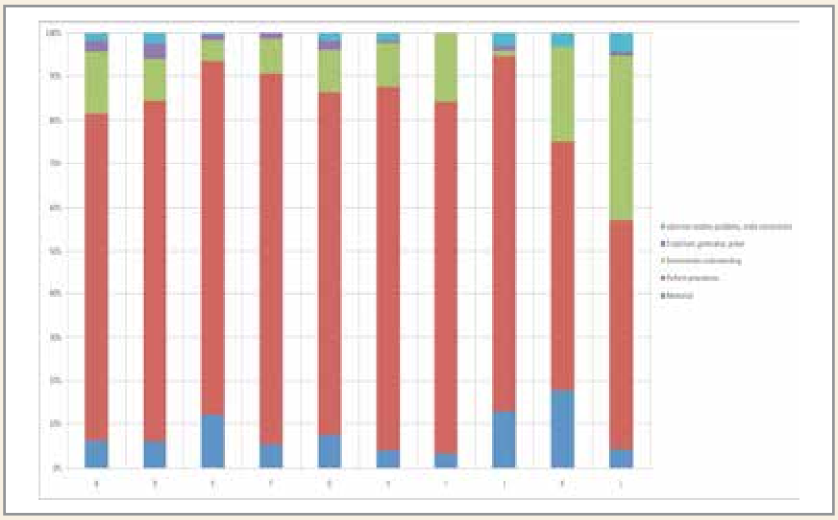


Table 2 shows the proportion of cognitive operations assessed in the selected Mathematics examinations. This mapping category was a key indicator of the design and profile of assessments and in Mathematics there are a number of interesting findings. The five mathematical cognitive operations are descriptions of the types of tasks or processes that mathematics educators have agreed should be evident in assessments (more expanded detail on the types of mathematical procedures subsumed into each cognitive operation can be found on page 4). There is an implied hierarchical order in these operations, which will be explored further in this paper. However, there are some marked similarities and differences across areas B–L. Performing mathematical procedures dominates assessments, ranging from 55% (L) to 80% (J) of coverage. Although small as an overall percentage, assessment items using memory skills also vary considerably across areas, ranging from 3% (I) to 18% (K). The variance of items that assess mathematical understanding is more marked, ranging from 2% (J) to 40% (L).

Those cognitive operations that should employ higher-order thinking skills (conjecture, generalise, prove' and 'solve non-routine problems, make connections') were represented very little and in some areas, not at all. For example Area I had no items assessing the last two cognitive operations. 'Conjecture, generalise, prove' only ever reached 4% coverage (D) and solving non-routine problems and making connections 4% (L). B and D were comparatively well represented in these higher-order cognitive operations, however it can be seen internationally that the general lack of their representation does need consideration when designing mathematics assessments that reflect all of the cognitive operations that constitute desired mathematical competencies.

Table 3: Cognitive demand

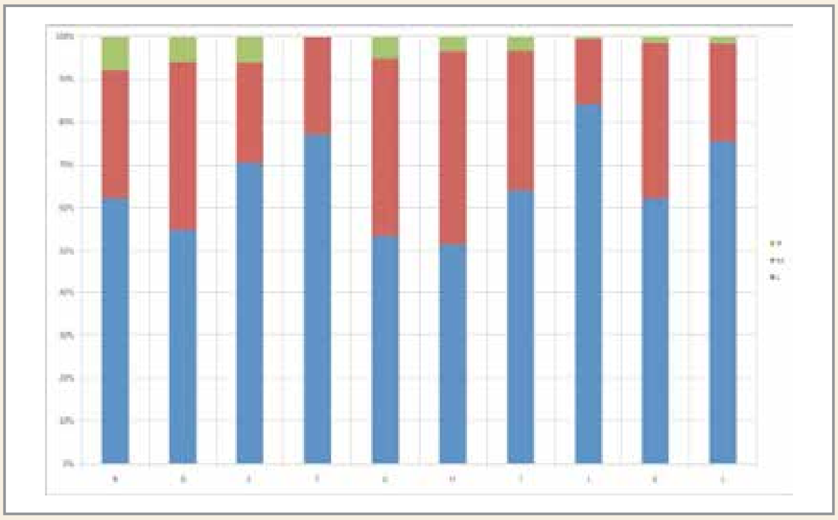


Table 3 shows the profile of assessment items of the basis of cognitive demand. Alongside cognitive operations, this category was a key indicator of the design and profile of assessments. More detail on how the cognitive demand criteria were established and operationalised is given on pages 5–8, however in short, this category describes the interrelation between the complexity of assessment items, and the expectations of their associated mark schemes and content standards. In general, low demand items dominated the assessments, with a range between 50% (H) to 83%(F). There were three main reasons for this; mostly the items involved straightforward one-step procedures, **or** the mark schemes were constructed atomistically so that there was no requirement for any linked steps to gain credit **or** the content standards were so specifically described that items required no application, just repetition of a routine procedure. Items of a medium cognitive demand varied significantly in proportion, ranging from 17%(J) to 48% (H), while highly cognitively demanding items were a rarity, ranging from 0% (F) to 9% (B). This table says a lot about how Mathematics is operationalised, and clearly has linkages to Table 2. (B) and (D) had the highest proportion of highly cognitive demanding items.

Table 4: Question type

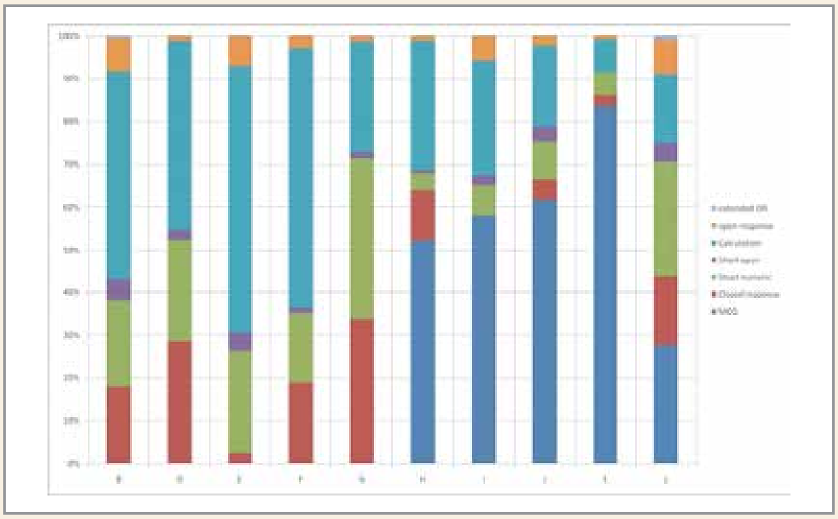


Table 4 shows how different examination boards and jurisdictions make use of different question types in Mathematics. It can be seen that there is considerable variation. Some international areas make significant use of multiple-choice questions (MCQs), particularly (K), while (B) made no use of these question types. The choice of assessment type is often associated with national or regional assessment systems. Apart from MCQs, unsurprisingly, short numeric and calculations were the most common question types, but it can also be seen that some areas, they make use of open written responses, usually in the form of justifying approaches. The choice of question types in their own right does not indicate much, however, Table 6 and 7 illustrate how successful question types were in assessing cognitive operations and cognitive demand.

Table 5: Cognitive operation v cognitive demand

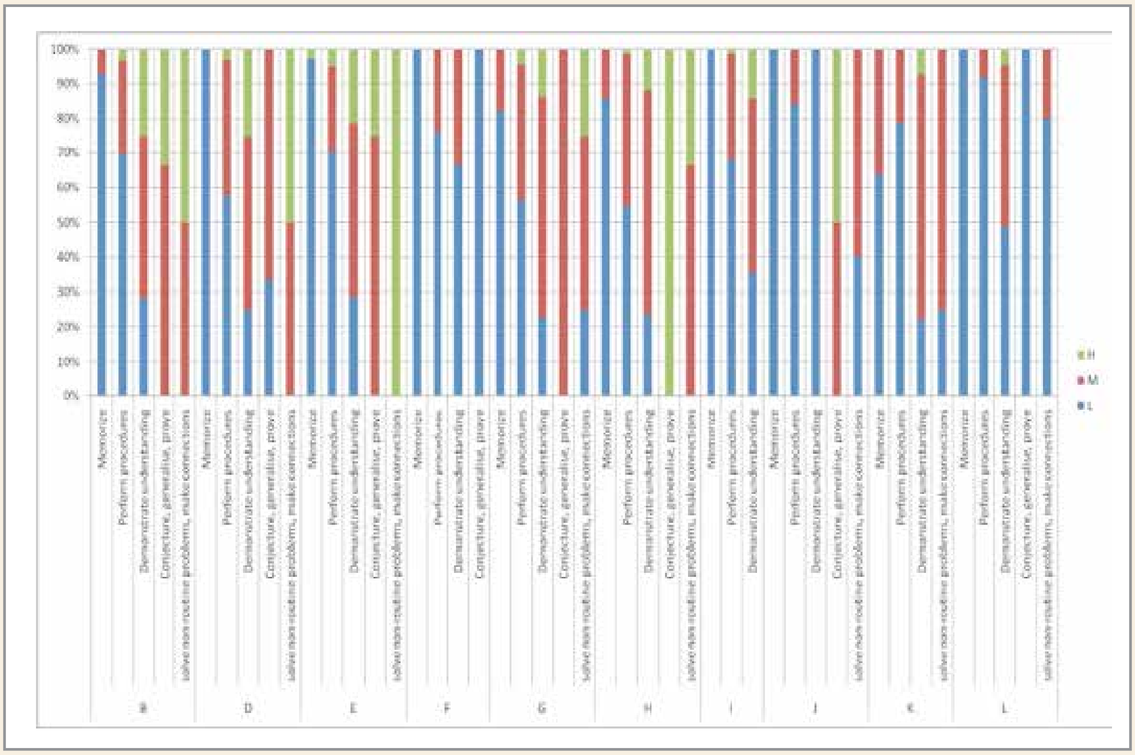
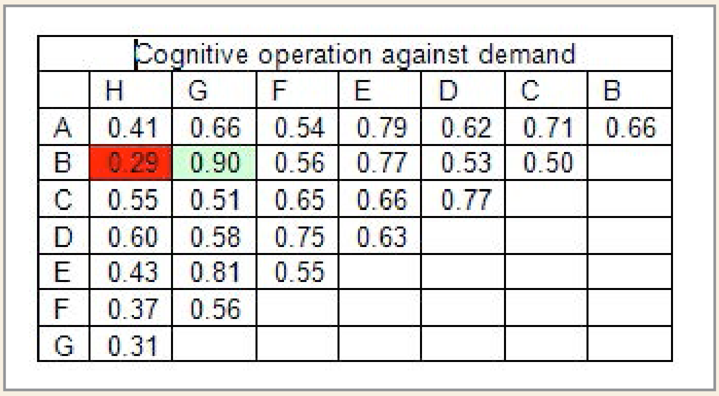


Table 5 shows the first of the correlations between factors, and is significant in terms of the assumption that the type of assessment task (cognitive operation) directly indicates the cognitive demand. It might be assumed that mathematics assessment items would become cognitively more demanding as cognitive operations move from those requiring memory skills across to those involving solving non-routine problems and making connections. Table 5 establishes this as a simplistic assumption. In almost every cognitive operation category across all examination boards/jurisdictions, the level of demand varied, according to the type of structure in which a question was presented, the expectations of the mark schemes and the manner in which the content standards had been written in terms of their specificity and transference to assessment items. If there are working assumptions that cognitive operations and cognitive demand should be aligned in terms of the design of assessments, it can be seen in Table 5 that some assessment structures are more effective than others in terms of the gradation of low to high demand items across the cognitive operations.

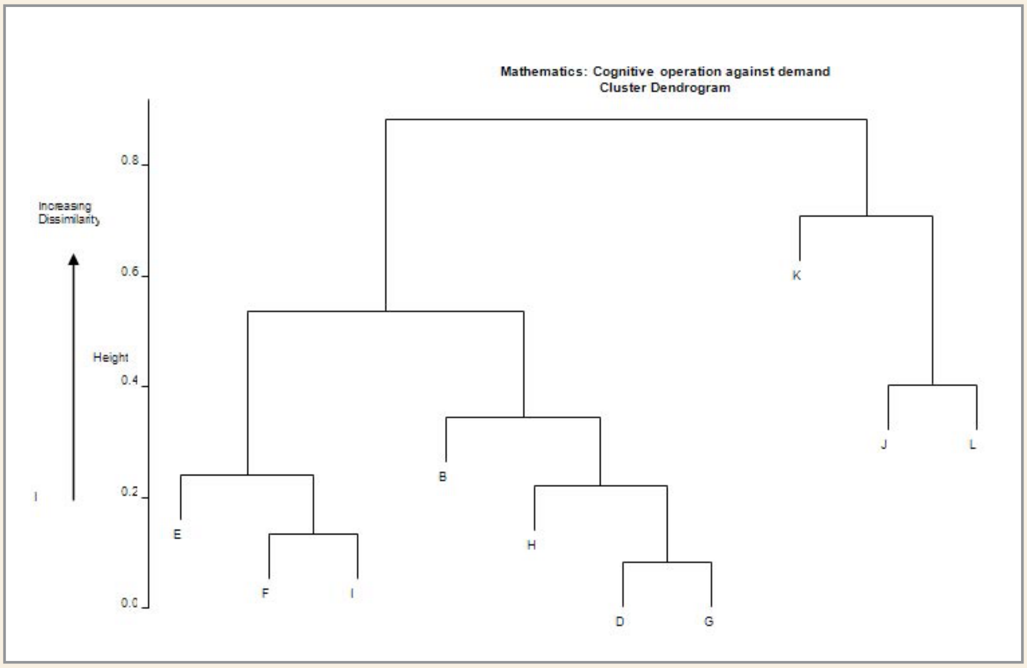
Table 6: Cognitive operation against demand



Cognitive operation against demand							
	H	G	F	E	D	C	B
A	0.41	0.66	0.54	0.79	0.62	0.71	0.66
B	0.29	0.90	0.56	0.77	0.53	0.50	
C	0.55	0.51	0.65	0.66	0.77		
D	0.60	0.58	0.75	0.63			
E	0.43	0.81	0.55				
F	0.37	0.56					
G	0.31						

The highlighted cells indicate the highest and lowest alignments.

Figure 3: Cognitive operation against demand dendrogram



The mathematics dendrogram (Figure 3) visually represents the tables of alignment indices. It shows the similarities between the relationship between cognitive operations and cognitive demands across jurisdictions. Close similarities are shown as clusters. The clusters at the bottom of the dendrogram show the most alignment, and the height between clusters shows the relative differences. Horizontal distances do not convey meaning.

The dendrogram indicates three paradigmic families of assessments. The most closely aligned assessment in terms of how they assessed cognitive operations and cognitive demand were (D) and (G) – which were then related to (H) and (B). (I) was closely aligned to (F) and then to (E). Interestingly, on the right-hand side, (L), (J) and more remotely (K) are related and all use psychometrics in their construction of their assessments.

Whether a position in paradigmic family groupings is seen as good or bad depends of course on the characteristics of the associated family members. The constituent factors of cognitive operations and cognitive demand are key to good assessments. (D) had relatively good proportions of cognitive operations and also cognitive demand and therefore **that** assessment served as a positive comparative assessment to be aligned with. Therefore similarities have to be based on whether it is considered good or bad to have the constituent profiles of the chosen features.

Table 7: Cognitive operation v question type

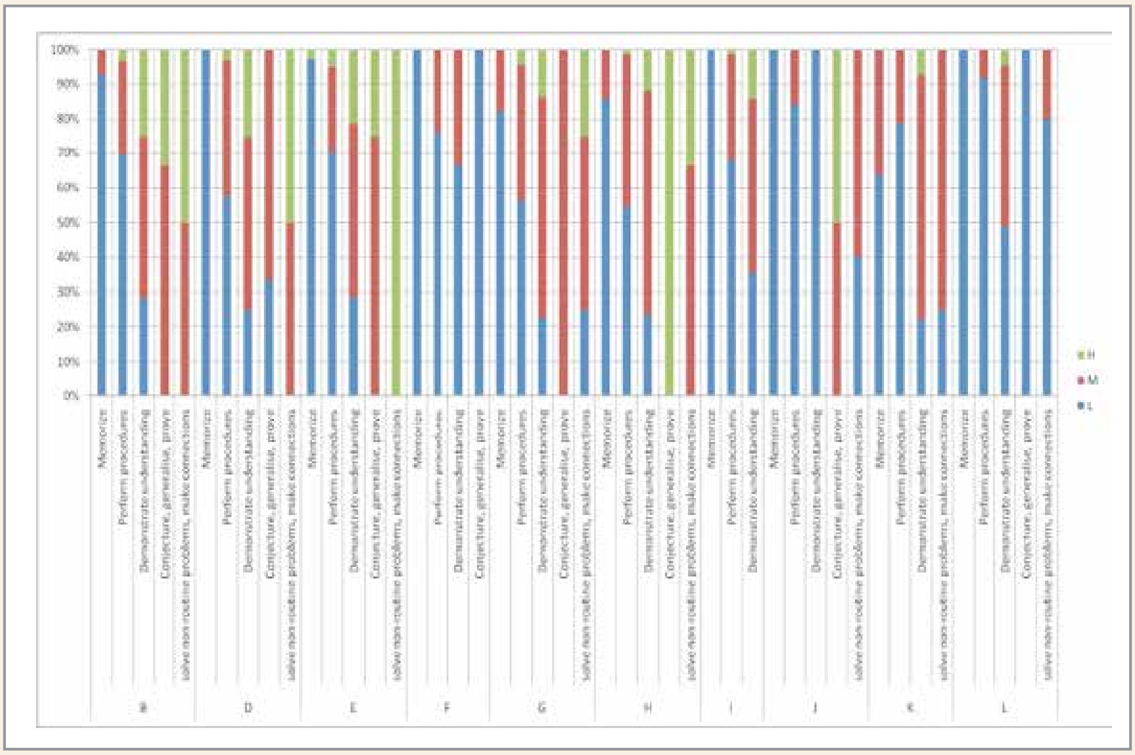
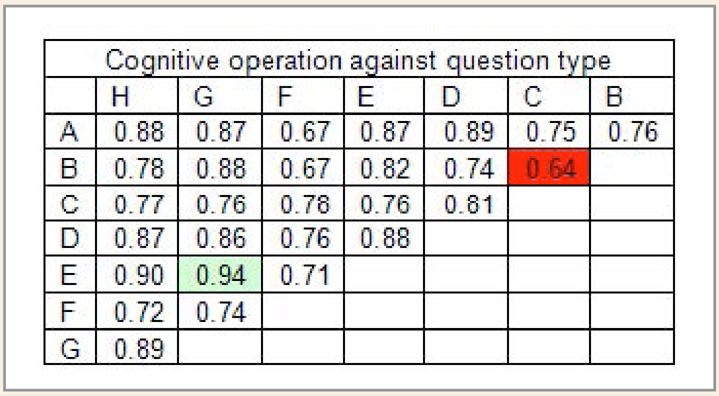


Table 8: Cognitive operation against question type



Cognitive operation against question type							
	H	G	F	E	D	C	B
A	0.88	0.87	0.67	0.87	0.89	0.75	0.76
B	0.78	0.88	0.67	0.82	0.74	0.64	
C	0.77	0.76	0.78	0.76	0.81		
D	0.87	0.86	0.76	0.88			
E	0.90	0.94	0.71				
F	0.72	0.74					
G	0.89						

Figure 4: Cognitive operation against question type dendrogram

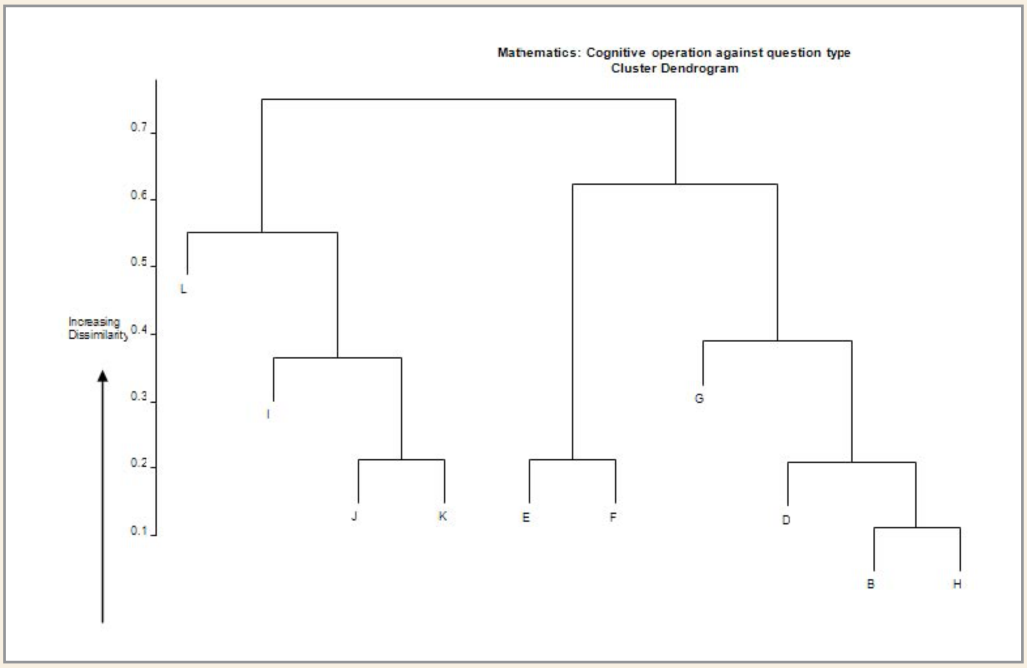


Table 9: Cognitive operation v question type

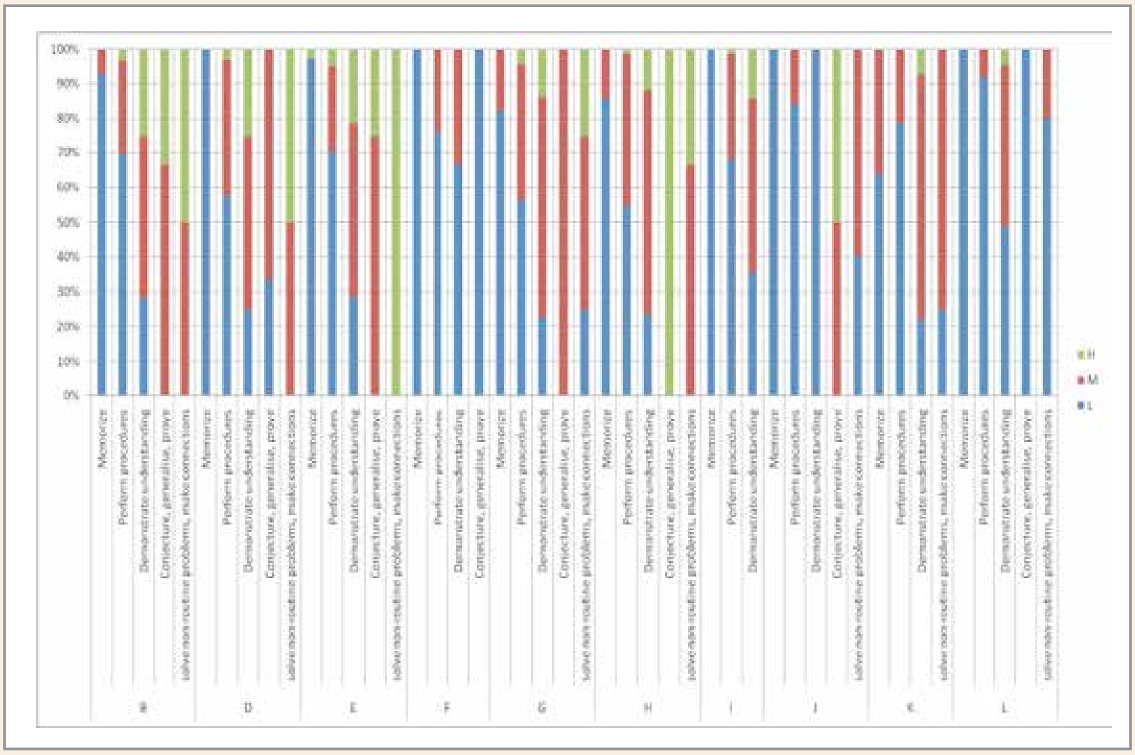
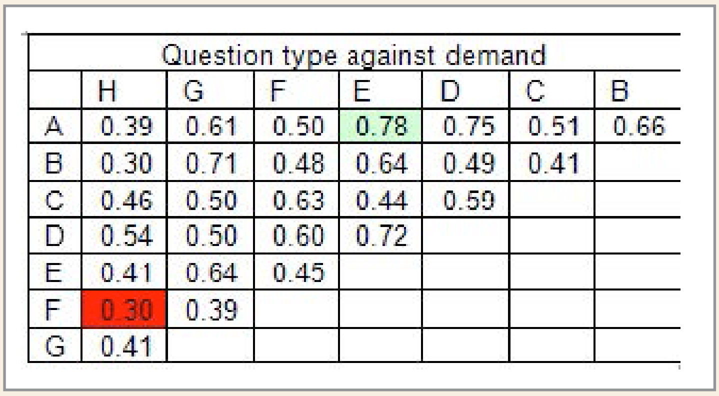


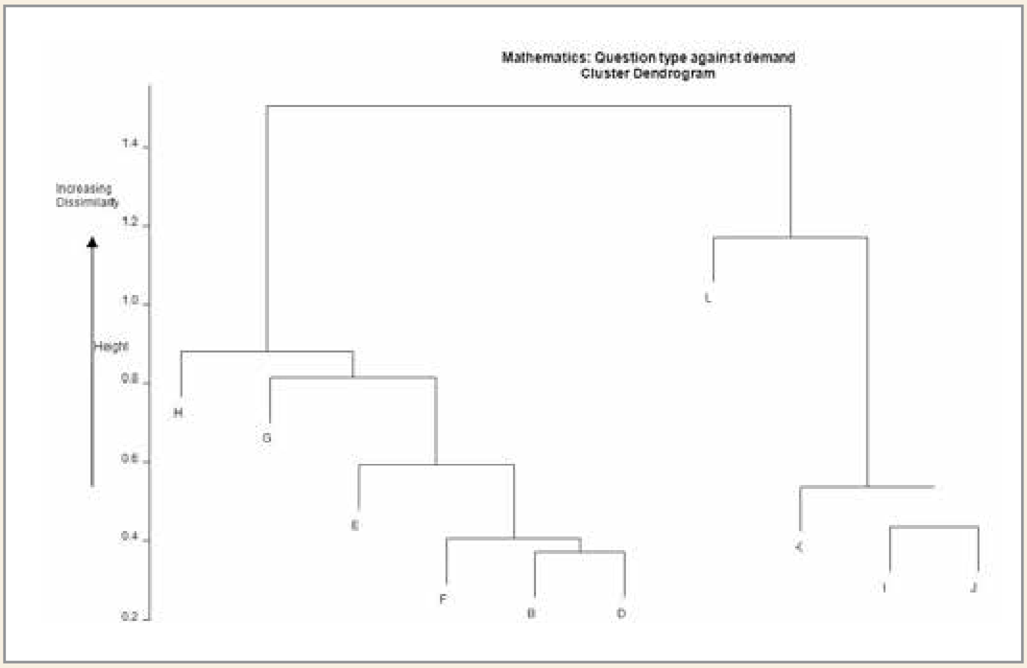
Table 10: Question type against demand



Question type against demand

	H	G	F	E	D	C	B
A	0.39	0.61	0.50	0.78	0.75	0.51	0.66
B	0.30	0.71	0.48	0.64	0.49	0.41	
C	0.46	0.50	0.63	0.44	0.59		
D	0.54	0.50	0.60	0.72			
E	0.41	0.64	0.45				
F	0.30	0.39					
G	0.41						

Figure 5: Question type against demand dendrogram



Tables above data shows profiles of how types of assessment tasks or processes students are asked to perform and their cognitive demand were linked to the question types used. Although there is variety both within and across areas, there were a few observed general trends. There are a few areas that did not use MCQ question types at all, however they formed a significant proportion of some international assessments. In general, MCQ, closed responses and short numeric question types were used to assess the lower-order cognitive operations. Similarly, although all question types can be used to assess all levels of cognitive demand, there is more likelihood of assessing higher-order cognitive skills through more open question types. The alignment tables and dendrograms were remarkably similar for all the correlatons in terms of paradigmic clusters in mathematics assessments. There is an indication that these paradigms have a relationship to the manner in which these jurisdictions develop assessment instruments, i.e. from psychometric or construct validity-based perspectives.

Key findings for Mathematics

The key comparative features of mathematics assessments across the selected examination boards and jurisdictions were:

- The uniform content standards focused on 'number', 'algebra' and 'geometry' in large but in varying proportions. Number ranged from 13% (H) to 29% (L), algebra ranging from 22% (J) to 46% (F) and geometry ranging from 21% (B) to 32% (J).
- 'Measure', 'probability' and 'statistics' were generally minor assessment areas, and in some cases not assessed at all. Measure ranged from 1% (I) to 10%.
- (G), probability ranging from 0% (F, I and K) to 10% (D) and statistics ranging from 1% (K) to 21% (L). (G) covered all content areas in the most even proportions.
- Sampling strategies, if used, were significantly different across jurisdictions in terms of the emphasis of mathematical content areas.
- Performing mathematical procedures dominates assessments, ranging from 55% (L) to 80% (J) of coverage.
- Although small as an overall percentage, assessment items using memory skills also vary considerably across areas, ranging from 3% (I) to 18% (K).
- The variance of items that assess 'mathematical understanding' is more marked, ranging from 2% (J) to 40% (L).
- Those cognitive operations that should employ higher-order thinking skills ('conjecture, generalise, prove' and 'solve non-routine problems, make connections') are represented very little and in some areas, not at all. For example Area I had no items assessing the last two cognitive operations.
- 'Conjecture, generalise, prove' only ever reached 4% coverage (D) and solving non-routine problems and making connections 4% (L).
- (B) and (D) were comparatively well represented in these higher-order cognitive operations, however it can be seen internationally that their general lack of representation does need consideration when designing mathematics assessments that reflect all of the cognitive operations that constitute desired mathematical competencies.
- In general, low-demand items dominate the assessments, with a range between 50% (H) to 83%(F).
- Items of a medium cognitive demand varied significantly in proportion, ranging from 17% (J) to 48% (H) of assessments.
- Highly cognitively demanding items were a rarity, ranging from 0% (F) to 9% (B).

The alignment indices, using Porter's Alignment Index equation, applied to aspects of assessments themselves rather than between content and assessment, show how closely all the jurisdictions are related to each other in terms of the correlated analyses and the accompanying dendrograms visually illustrate these relationships in terms of clusters or families of similar features. It can be seen that on the basis of this international comparative assessment research, there are significant similarities and differences between groups of assessments. In particular, the relationship between how cognitive operations and cognitive demand are assessed is of interest. In conjunction with the bar chart analyses, it has been possible to explore at item level the underlying reasons for similarities and differences between assessments, and be informed by good or desired practice. On the basis of the best ranges of cognitive operations and cognitive demands, (G), (H), (B) and (D) served as the most positive comparative assessments to be aligned with, with the caveat that there could be significantly better coverage, if designed for. Overall, a key message has been that good mathematics assessments do not happen by accident: they require active assessment design followed by good assessment writing. Having dissected national and international assessments in some detail, desired assessment principles and outputs can be developed and sustained using robust classification rubrics for all the required conditions of assessments, sampled content representation, a range of cognitive operations, identifiable cognitive demands and a range of question types purposed appropriately.

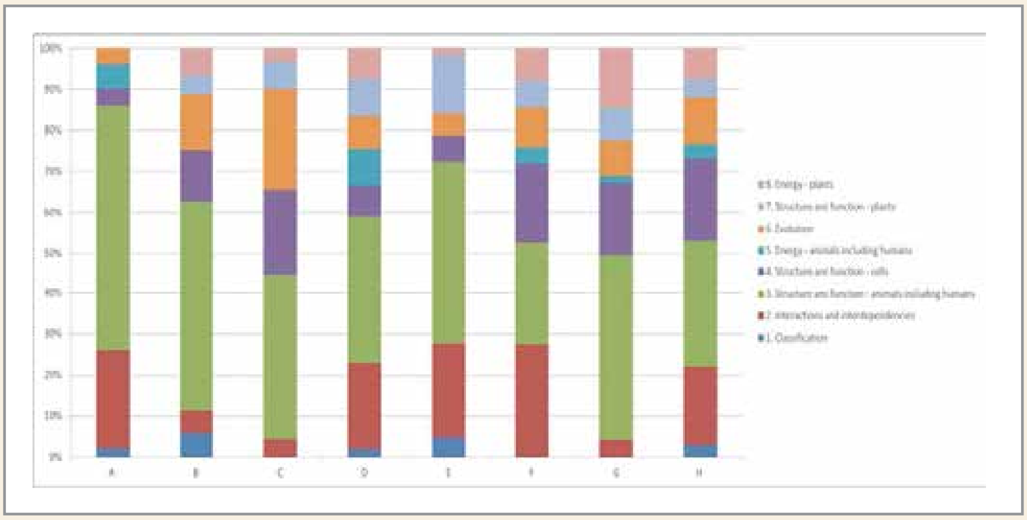
A criticism of some Mathematics curricula is the 'mile wide and inch deep' stigma. This inference has often been levelled as a reason that certain jurisdictions perform poorly in Mathematics in the content assessments of TIMSS. This research suggests that this criticism is not well founded in assessment evidence. The jurisdictions that had significantly less content representation in terms of mathematical areas did not show evidence of assessing a wider range of cognitive operations or at higher cognitive demands. These factors seem to work more independently, and need to be actively designed into assessments. It was of interest to note that the cognitive operations that are often cited as those to develop higher-order and transferable thinking skills were significant in their scarcity across jurisdictions. If desired, these skills and assessment items need to be focused on, promoted in terms of content standards and then represented appropriately in assessments.

Mapping outcomes for the sciences

The results of the mapping analyses for Biology, Physics, Chemistry and Science are illustrated and discussed in this section. As described earlier, there were fewer comparative international jurisdictions for the separate sciences than there were for Mathematics and 'science'.

Biology international comparisons

Table 11: Content standards



It can be seen from Table 11 that there was considerable variation on proportions of questions from the eight Biology content areas across national and international assessments. Similarly to the discussion surrounding the coverage of Mathematics content standards, linear assessments usually use sampling strategies over time to ensure content representation. As this study was purposed to look across a number of examination boards and jurisdictions over one or two examination series, the content coverage may not be generalisable. However, on the basis of this evidence, there were significant differences in the emphases of different aspects of biology assessed in examinations.

The assessment content standard that dominated the selected biology assessments was 'the structure and function of animals'. This being said, coverage across countries varied from approximately 25% (F) to 60% (A) of any examination. Assessment items on 'interactions and dependencies' were relatively well represented with a few exceptions, as were items related to 'cell biology' and 'evolution'. It can also be seen that some biological areas are not assessed in some jurisdictions, e.g. in 'classification', (C), (F) and (F) , the 'structure and function of plants' and 'energy in plants' (A). There were significant differences in assessment emphasis across examination boards and jurisdictions of content areas. Whether these differences are by design or a result of some form of random sampling would need to be established before any conclusions made about under or over representation. Similarly for mathematics assessments, how differing examination boards or jurisdictions create or maintain a sampling strategy with an associated justification and rationale is a key assessment issue in terms of validity, reliability, standard-setting and maintenance.

Table 12: Cognitive operations

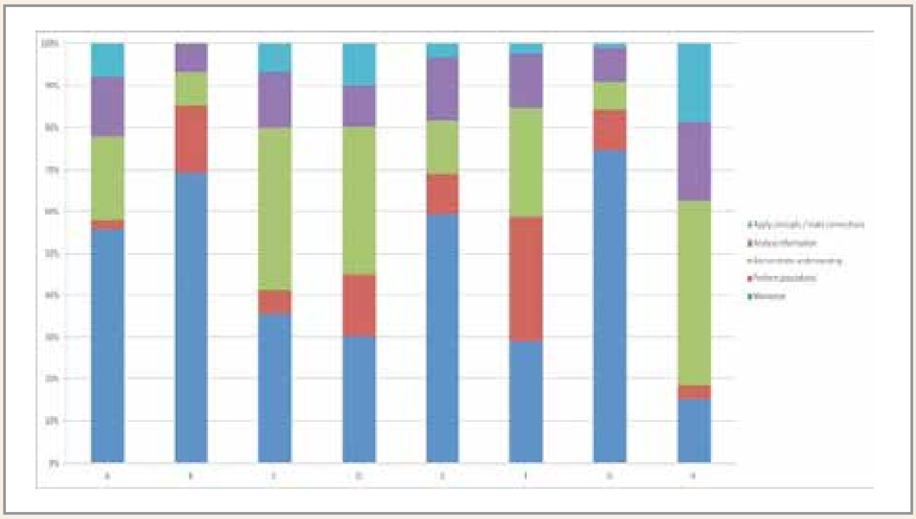


Table 12 shows the proportion of cognitive operations assessed in the selected biology examinations. This category was a key indicator of the design and profile of Biology assessments. The five biology cognitive operations are descriptions of the types of tasks or processes that biology examiners and practitioners have agreed should be evident in assessments (more expanded detail on the types of biology procedures subsumed into each cognitive operation can be found on pages 4–5). The striking finding is the high proportion of assessment items requiring memory and recall. Over 75% of jurisdiction (G) used this cognitive operation, while jurisdiction (H) used memory the least with approximately 25%. Another interesting cognitive operation comparison is 'performing procedures'. In the sciences, this is normally associated with enquiry-based skills – a desired focus in terms of curriculum and assessment. Looking at how much this cognitive operation is represented in examinations is interesting, from jurisdiction (F) having a significant proportion (30%), to just 1% in jurisdiction (A). A key question to ask is where this cognitive operation is represented; through classroom practice, teacher-assessed investigative components, a practical exam or a component of an exam? Most science educators value its essential role in science education (e.g. Millar and Osborne, 1998), however it needs an established space or else it risks becoming marginalised in emphasis. Items requiring demonstrating understanding also varied significantly across examination boards and jurisdictions, ranging from 5% (G) to 45% (H), while those cognitive operations that should employ higher-order thinking skills, analysing information and applying concepts and making connections were represented very little (e.g. G), and in some areas, not at all (B). (H) had the most even spread of cognitive operations, except for performing operations.

Table 13: Cognitive demand

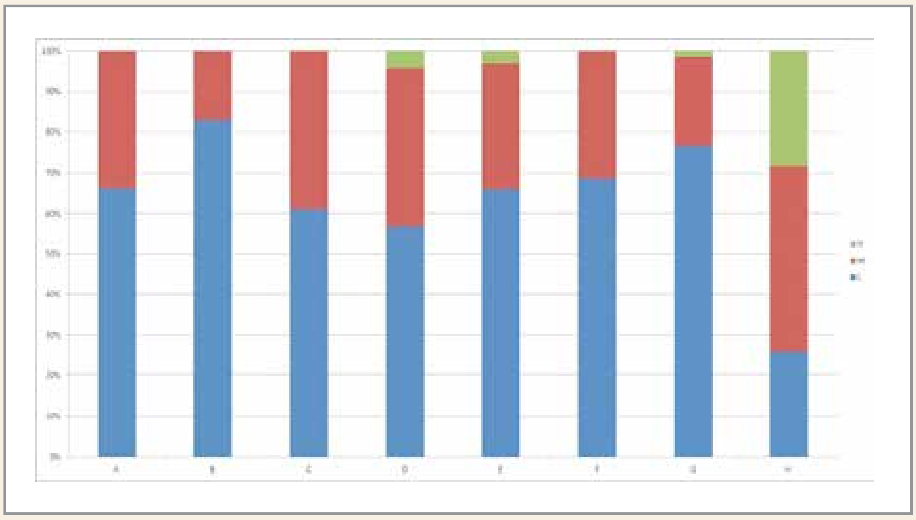


Table 13 shows the profile of assessment items of the basis of cognitive demand. Alongside cognitive operations, this category was a key indicator of the design and profile of assessments. More detail on how the cognitive demand criteria were established and operationalised is given on pages 5–8 however in short, this category describes the interrelation between the complexity of assessment items, and the expectations of their associated mark schemes and content standards. It is clear, that apart from (H), low cognitive demand items dominated the assessments, ranging from 26% (H) to 82% (B). Items of medium demand items were reasonably represented, ranging from 20% (G) to 45% (H). Highly cognitively demanding items were a rarity, hardly registering at all across jurisdictions, apart from(H) with an impressive 29%. This table says a lot about how biology assessments are operationalised, and clearly has linkages to Table 12.

Table 14: Question type

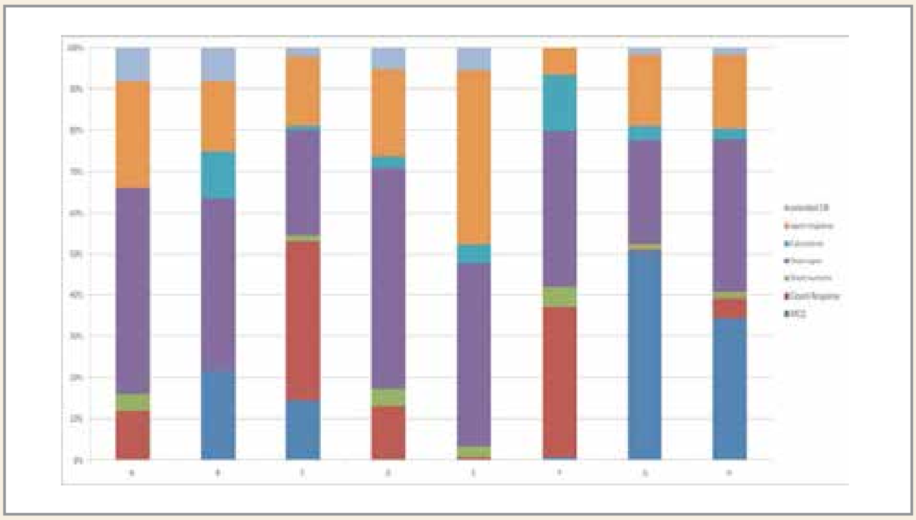


Table 14 shows how different examination boards and jurisdictions made use of different question types in Biology. It can be seen that there was considerable variation. Some international areas made significant use of multiple-choice questions (MCQs), particularly (G), while others made little or no use of these question types (e.g. F and A). Apart from MCQs, there were significant proportions of short and longer open-response items. It is of interest to compare the use of items requiring calculations, ranging from 0% (A) to15% (F). The under-representation of mathematics in science assessments is an issue, and while the content representation of these assessments may explain some of the absence of calculations, it is an identifiable gap in a number of jurisdictions.

Table 15: Cognitive operation v cognitive demand

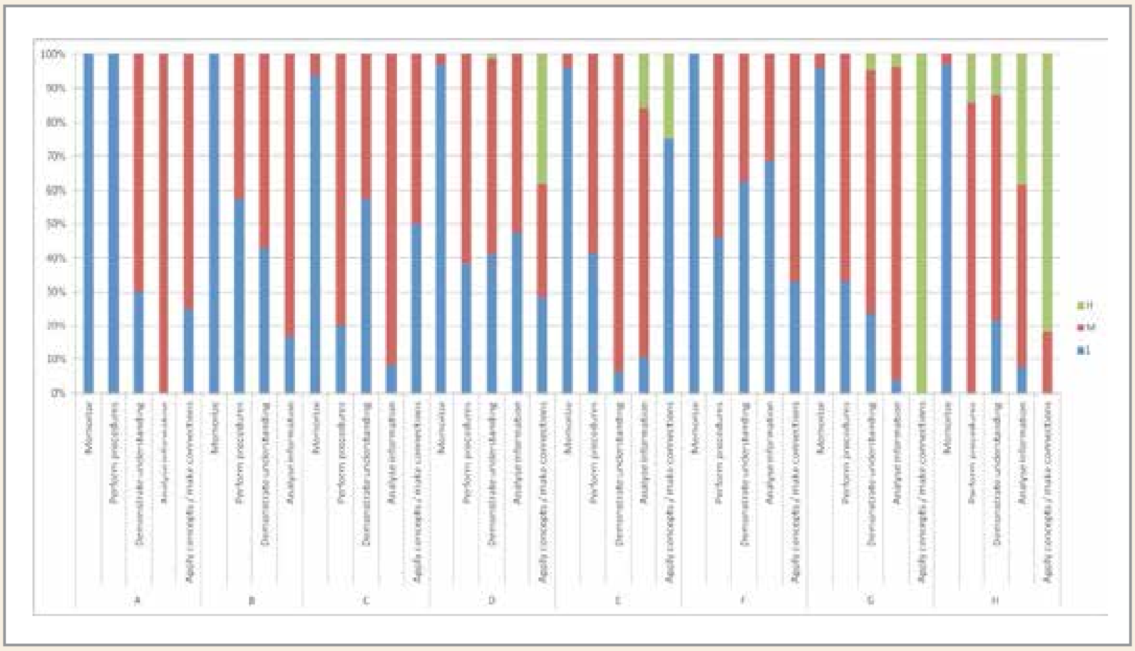
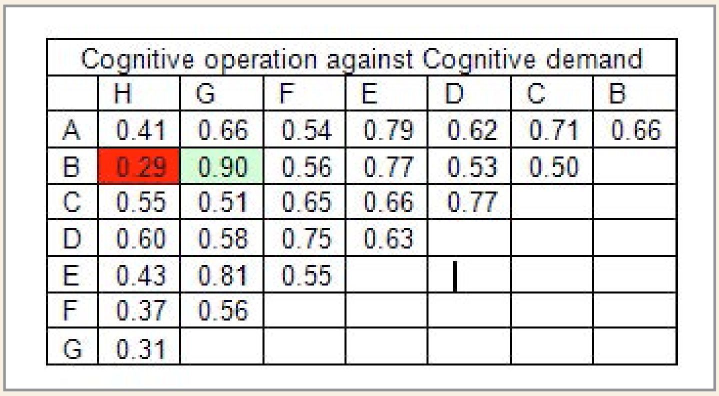


Table 16: Cognitive operation against cognitive demand



Cognitive operation against Cognitive demand							
	H	G	F	E	D	C	B
A	0.41	0.66	0.54	0.79	0.62	0.71	0.66
B	0.29	0.90	0.56	0.77	0.53	0.50	
C	0.55	0.51	0.65	0.66	0.77		
D	0.60	0.58	0.75	0.63			
E	0.43	0.81	0.55				
F	0.37	0.56					
G	0.31						

Figure 6: Cognitive operation against demand dendrogram

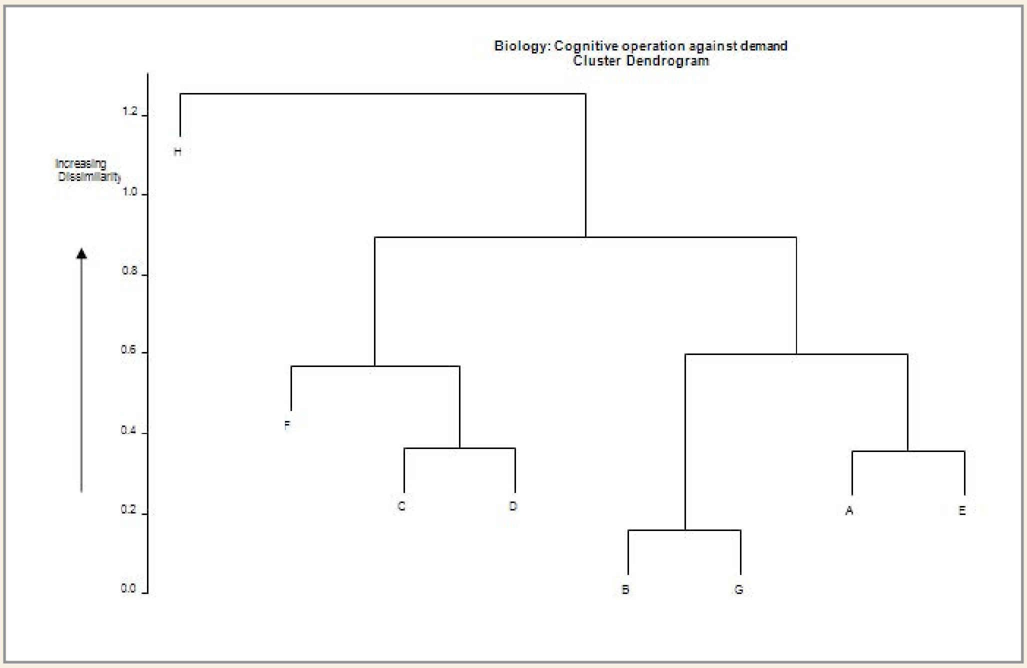


Table 15 shows the correlation between the cognitive operations (the types of tasks or processes) and the cognitive demand of biology assessments. As discussed in the previous section on mathematics, it might be assumed that biology assessment items would become cognitively more demanding as cognitive operations move from those requiring memory skills across to those involving solving non-routine problems and making connections. Table 15 does indicate that memory items are almost entirely of low demand (and Table 12 has already established that memory/recall-related items dominate biology assessments). Table 15 also indicates that cognitive demand cannot be assumed by the type of assessment task. Although high-demand questions (although very few in number) were more often found in applying concepts and making connections, most examination boards and jurisdictions presented a mixed profile of types of tasks to their cognitive demand.

The biology dendrogram (Figure 6) visually represents the tables of alignment indices. It shows the similarities between the relationship between cognitive operations and cognitive demands across jurisdictions. Close similarities are shown as clusters. The clusters at the bottom of the dendrogram show the most alignment, and the height between clusters show the relative differences. Horizontal distances do not convey meaning.

The dendrogram indicates three main paradigmic families of assessments. The most closely aligned assessments in terms of how they assessed cognitive operations and cognitive demand were (B) and (G), which had some relationship with (A) and (E). In a separate cluster, (D) was aligned with (C) and then to (F). (H) had characteristics that were significantly different to all the others.

Whether a position in paradigmic family groupings is seen as good or bad depends of course on the characteristics of the associated family members. The constituent factors of cognitive operations and cognitive demand are key to good assessments. (H) had by far the best proportions of cognitive operations and cognitive demand, and clearly that assessment would be the most positive comparative assessment to be aligned with. However, as shown above, (H) could not be directly compared or aligned with any other jurisdiction. Judgements on alignments have to be based on whether it is considered good or bad to have the constituent profiles of the chosen features. In general this dendrogram suggests that the alignments of (B), (G), (A) and (E) are not particularly positive. The alignment between (D), (C) and (F) are more positive in terms of their assessment structures.

Table 17: Cognitive operation v question type

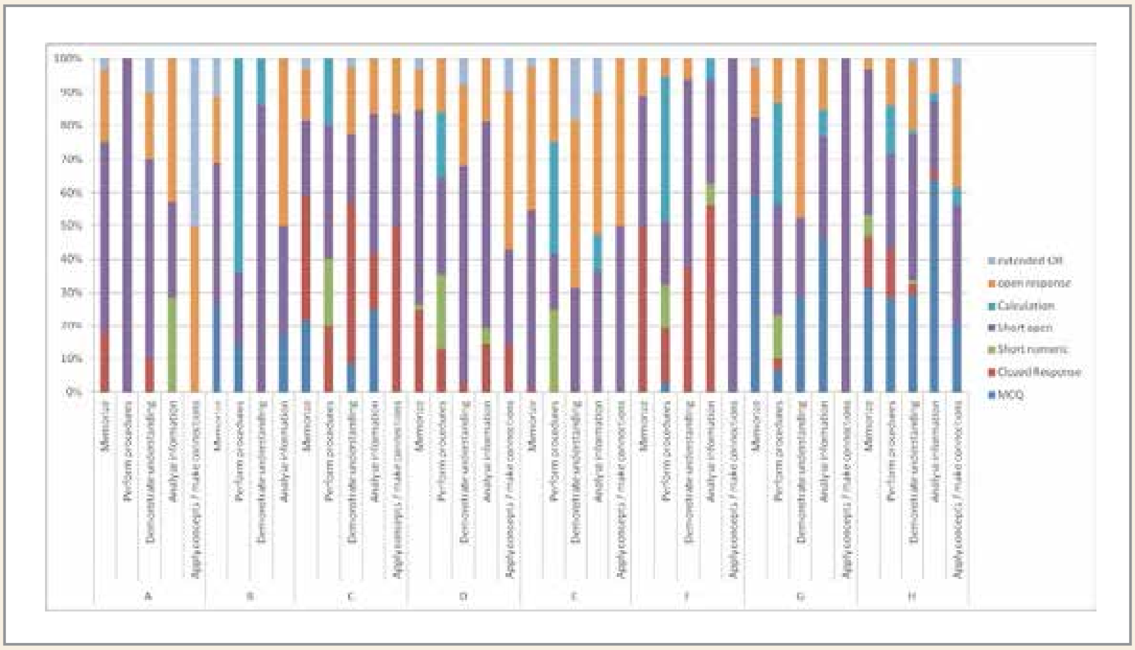
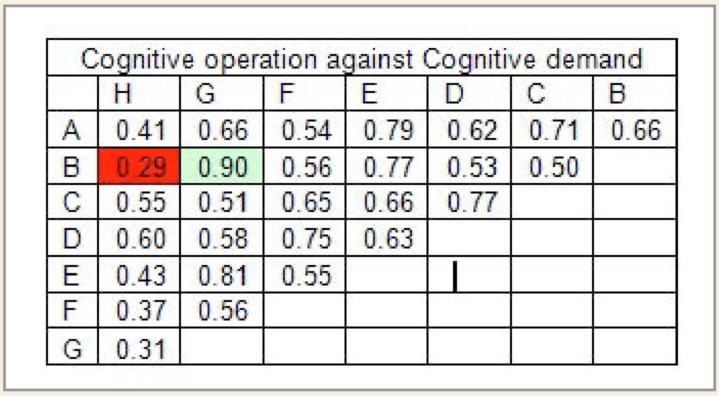


Table 18: Cognitive operation against question type



Cognitive operation against Cognitive demand							
	H	G	F	E	D	C	B
A	0.41	0.66	0.54	0.79	0.62	0.71	0.66
B	🟥 0.29	🟩 0.90	0.56	0.77	0.53	0.50	
C	0.55	0.51	0.65	0.66	0.77		
D	0.60	0.58	0.75	0.63			
E	0.43	0.81	0.55				
F	0.37	0.56					
G	0.31						

Figure 7: Cognitive operation against question type dendrogram

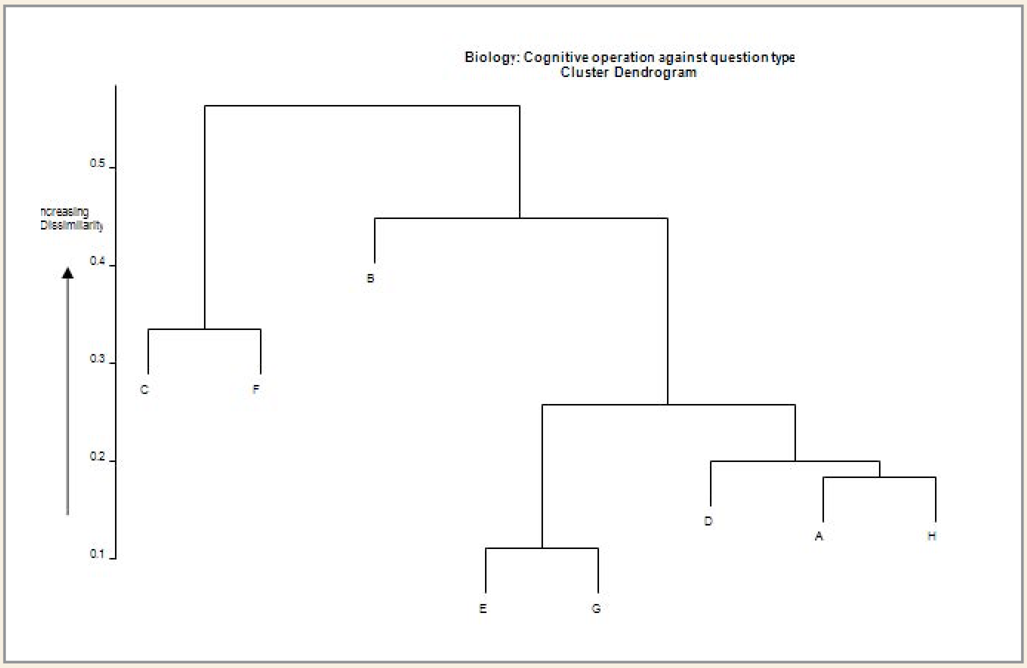


Table 19: Question type v cognitive demand

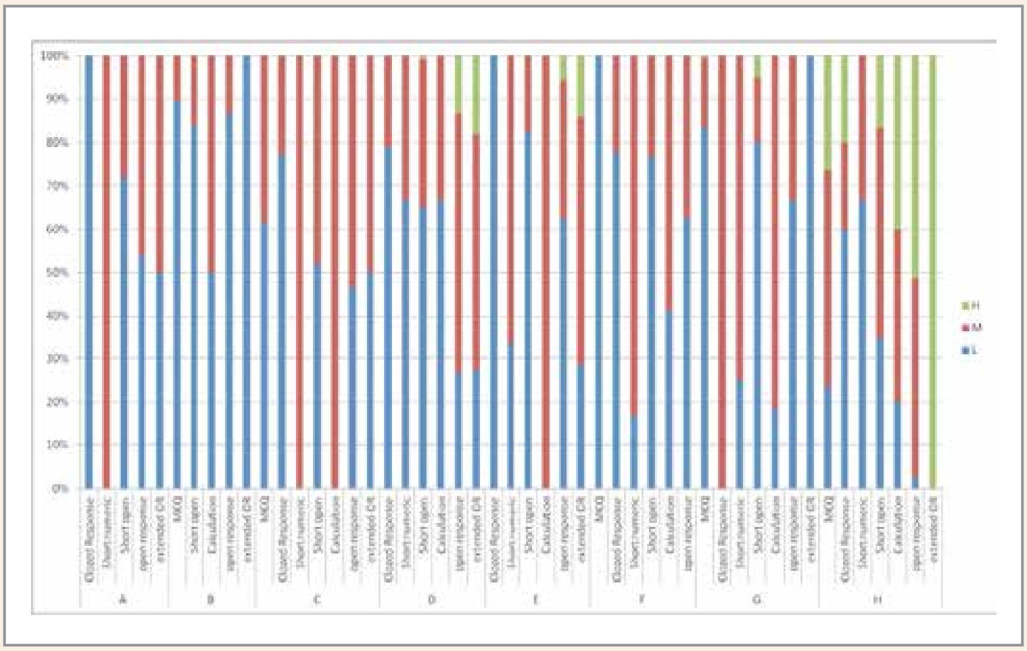
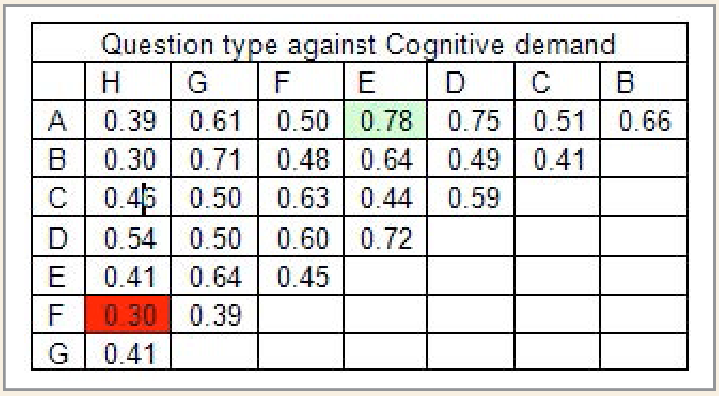
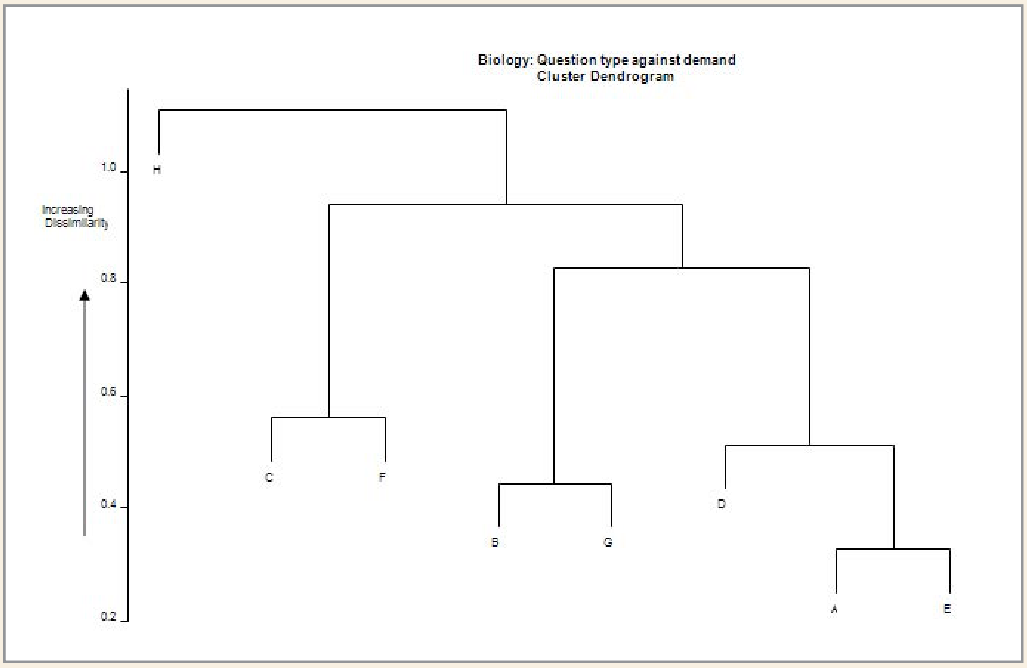


Table 20: Question type against cognitive demand



Question type against Cognitive demand							
	H	G	F	E	D	C	B
A	0.39	0.61	0.50	0.78	0.75	0.51	0.66
B	0.30	0.71	0.48	0.64	0.49	0.41	
C	0.46	0.50	0.63	0.44	0.59		
D	0.54	0.50	0.60	0.72			
E	0.41	0.64	0.45				
F	0.30	0.39					
G	0.41						

Figure 8: Question type against demand dendrogram



Tables 17 and 19 show profiles of how types of assessment tasks or processes students are asked to perform and their cognitive demand are linked to the question types used. These tables show that a broad range of question types is used to assess various cognitive operations, but apart from (H), open and more extended question types were no guarantee of the assessment of higher-order thinking skills. The dendrograms (Figures 6, 7 and 8) show different clustering arrangements on the basis of both of these conditions.

Key findings for Biology

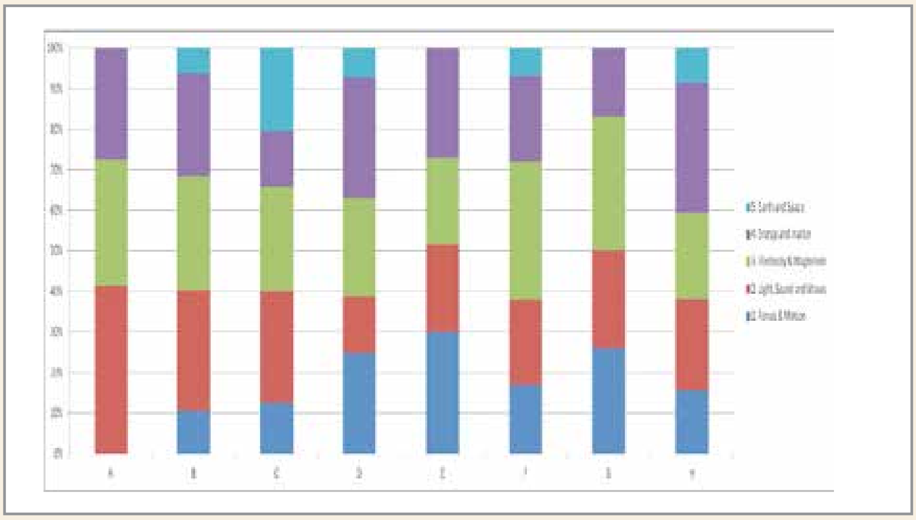
The key comparative features of biology assessments across the selected examination boards and jurisdictions were:

- The assessment content standard that dominated the selected biology assessments was 'the structure and function of animals'. This being said, coverage across countries varied from approximately 25% (F) to 60% (A) of any examination.
- Assessment items on interactions and dependencies were relatively well represented with a few exceptions, as were items related to 'cell biology' and 'evolution'.
- Some biological areas were not assessed at all, in particular classification, the structure and function of plants and energy in plants.
- Sampling strategies, if used, were significantly different across jurisdictions in terms of the emphasis of biology content areas.
- The cognitive operation of 'memorise' dominated the assessment of biology, but also itself had significant variation across jurisdictions, ranging from 25% in (H) to 75% (G).
- The cognitive operations requiring 'performing procedures' were characterised by scientific enquiry skills. Coverage of this cognitive operation varies from 1% (A) to 30% (F).
- The coverage of the cognitive operation of 'demonstrating understanding' varied significantly, ranging between 6% in (B) and 44% in (H).
- Higher-order cognitive operations ('analyse information and apply concepts/make connections') were in general assessed very little (G), and in some cases not at all (B).
- Low cognitive-demand items dominated the assessments (with significant variation), ranging from 26% (H) to 82% (B). There were significant proportions of medium-demand items, but again with significant variations across jurisdictions ranging from 20% (G) to 45% (H).
- In general, there were very few items at high demand (D, E, and G), and in many jurisdictions, none at all (A, B, C and F).
- The biology assessments from (H) were significantly different from the other jurisdictions. They had more balance in their cognitive operations and far more cognitive demand.

The alignment indices, using Porter's Alignment Index equation, were applied to aspects of assessments themselves (rather than the way Porter used them – between content standards and assessment) as the construction and purposes of content standards are not consistent internationally. The alignment indices showed how closely the jurisdictions were related to each other in terms of the three correlated analyses, and the accompanying dendrograms visually illustrate these relationships in terms of clusters or families of similar features. It can be seen that on the basis of this international comparative assessment research, there are significant similarities and differences between groups of assessments. In conjunction with the bar chart analyses, it has been possible to explore at item level the underlying reasons for similarities and differences between assessments, and be informed by identified good practice. On the basis of very good proportions of cognitive operations and cognitive demands (H) served as the most positive comparative assessments to be aligned with. Unfortunately, no other jurisdiction could be aligned with them. The dendrograms suggest that the alignments of B, G, A and E were not particularly positive. The alignments between D, C and F were more positive in terms of their assessment structures.

Physics international comparisons

Table 21: Content standards



It can be seen from Table 21 that there was considerable variation on proportions of questions from the five Physics content areas across national and international assessments. Linear assessments usually use sampling strategies over time to ensure content representation. As this study was purposed to look across a number of examination boards and jurisdictions over one or two examination series, the content coverage may not be generalisable. However, on the basis of this evidence, there were significant differences in the emphases of different aspects of Physics assessed in examinations.

Unlike Biology, no one content standard dominates the selected Physics assessments. This being said, coverage of content standards varied across jurisdictions and some Physics areas were not assessed at all. Whether these differences are by design or a result of some form of random sampling would need to be established before any conclusions could be made about under or over representation. How differing examination boards or jurisdictions create or maintain a sampling strategy with an associated justification and rationale is a key assessment issue in terms of validity, reliability, standard-setting and maintenance. It can be clearly seen that the assessments from (H) had the most evenly proportioned distributions between content areas.

Table 22: Cognitive operation

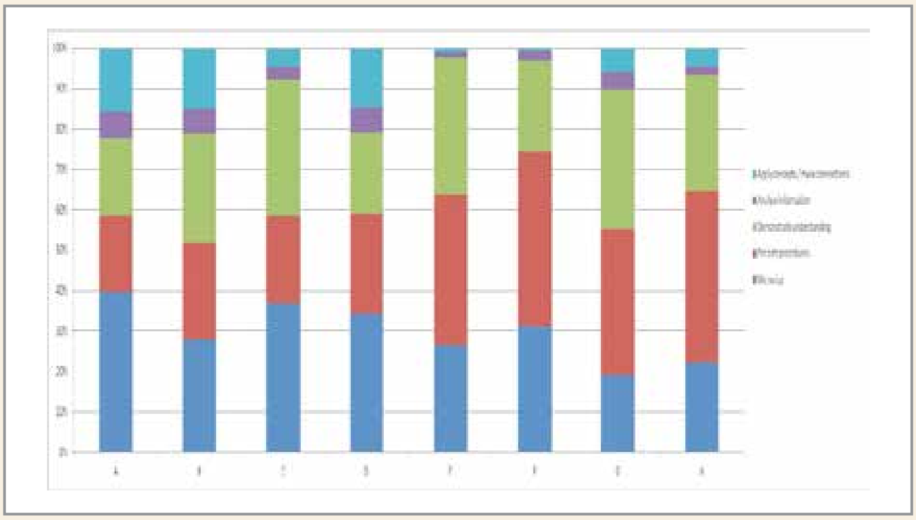


Table 22 shows the proportion of cognitive operations assessed in the selected physics examinations. This mapping category was a key indicator of the design and profile of assessments. The five physics cognitive operations are descriptions of the types of tasks or processes that physics examiners and practitioners have agreed should be evident in assessments (more expanded detail on the types of physics procedures subsumed into each cognitive operation can be found on pages 4–5, these are generic across all the sciences). The first three cognitive operations dominated the assessments across all the jurisdictions. Memorising ranged from 19% (G) to 39% (A), performing procedures, which is largely the assessment of enquiry skills, ranged from 19% (A) to 41% (H) and demonstrating understanding ranged from 19% (A) to 35% (G). Those cognitive operations that should employ higher-order thinking skills (analysing information and applying concepts and making connections) were generally represented very little, particularly in (E) at 3% and (F) at 4%. (A), (B) and (D) covered these and indeed all cognitive operations in relatively good proportions.

Table 23: Cognitive demand

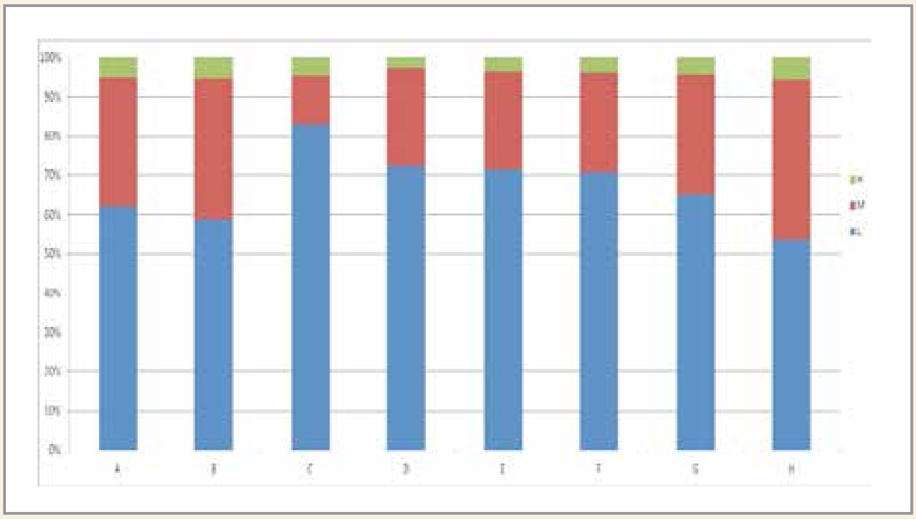


Table 23 shows the profile of assessment items of the basis of cognitive demand. More detail on how the cognitive demand criteria were established and operationalised is given on pages 5–8, however in short, this category describes the interrelation between the complexity of assessment items, and the expectations of their associated mark schemes and content standards. In general, low-demand items dominated the assessments, with a range between 53% (H) to 82% (C). There were three main reasons for this: mostly the items involved straightforward one-step procedures, or the mark schemes were constructed atomistically so that there was no requirement for any linked steps to gain credit **or** the content standards were so specifically described that items required no application, just repetition of a routine procedure. Items of a medium cognitive demand varied significantly in proportion, ranging from 14% (C) to 41% (H), while highly cognitively demanding items were generally few, ranging from 2% (D) to 6% (H).

Table 24: Question type

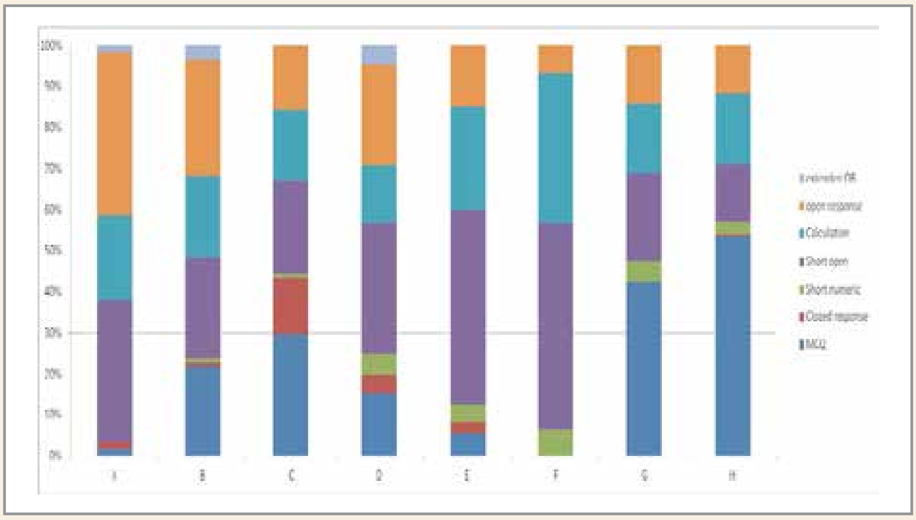


Table 24 shows how different examination boards and jurisdictions make use of different question types in physics. It can be seen that there was considerable variance. Some international areas made significant use of multiple-choice questions (MCQs) e.g. (H) and (G) while other areas do not use them at all, e.g. (F). Apart from MCQs, there were significant but varying proportions of short open and calculation response items. There was very little evidence of extended open-response items in physics questions.

Table 25: Cognitive operation vs cognitive demand

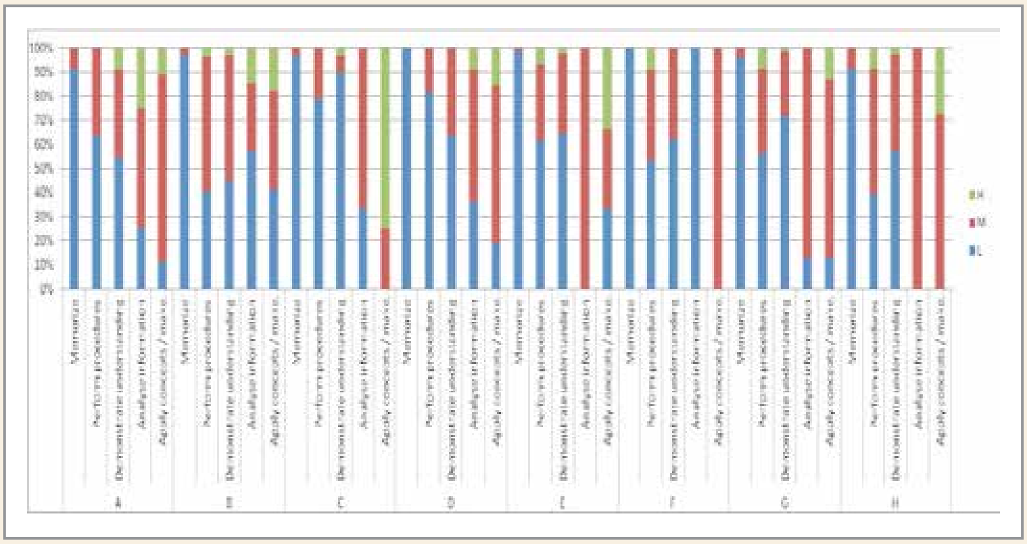
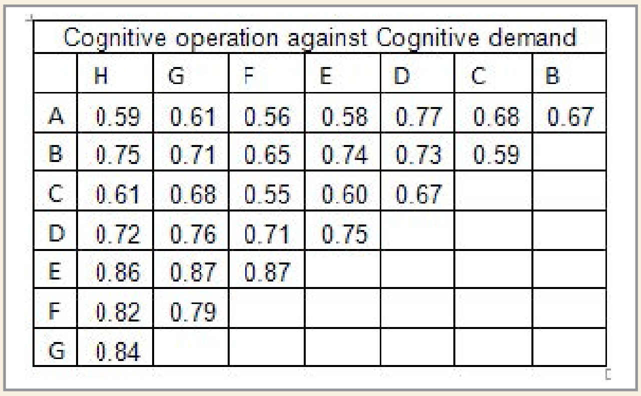


Table 26: Physics alignment indices



Cognitive operation against Cognitive demand							
	H	G	F	E	D	C	B
A	0.59	0.61	0.56	0.58	0.77	0.68	0.67
B	0.75	0.71	0.65	0.74	0.73	0.59	
C	0.61	0.68	0.55	0.60	0.67		
D	0.72	0.76	0.71	0.75			
E	0.86	0.87	0.87				
F	0.82	0.79					
G	0.84						

Figure 9: Cognitive operation against cognitive demand dendrogram

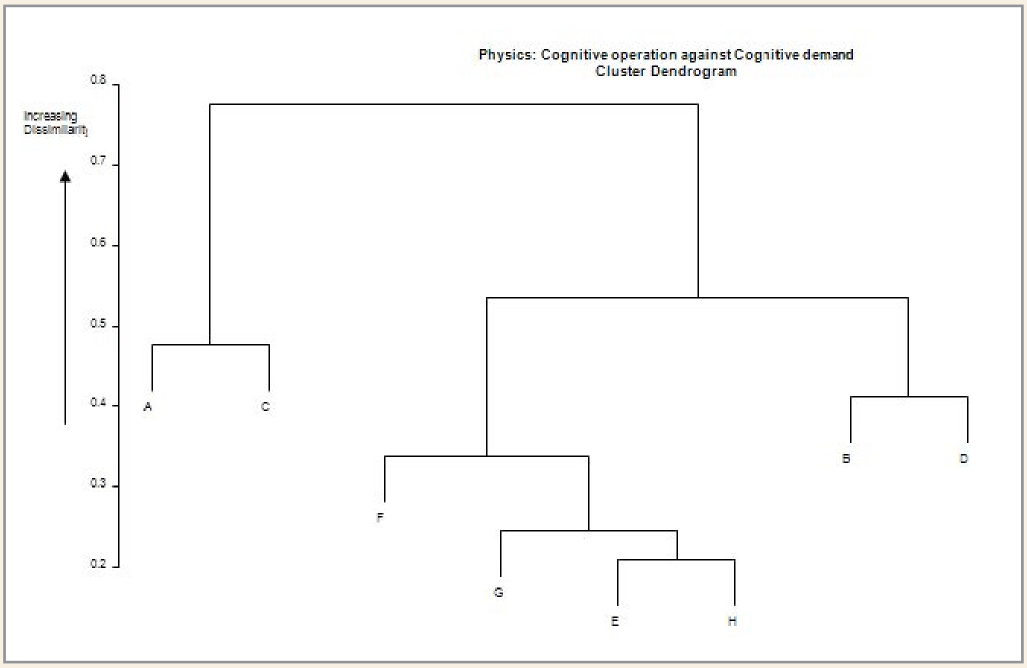


Table 25 shows the correlation between the cognitive operations (the types of tasks or processes) and the cognitive demand of physics assessments. It might be assumed that physics assessment items would become cognitively more demanding as cognitive operations move from those requiring memory skills across to those involving solving non-routine problems and making connections. Table 25 demonstrates this to be a simplistic assumption. Although memory items were almost entirely of low demand, Table 25 also indicates that cognitive demand cannot be assumed by the type of assessment task. Although high-demand questions (although few in number) were more often found in applying concepts and making connections, most examination boards and jurisdictions presented a mixed profile of types of tasks to their cognitive demand.

The physics dendrogram (Figure 9) visually represents the tables of alignment indices. It shows the similarities between the relationship between cognitive operations and cognitive demands across jurisdictions. Close similarities are shown as clusters. The clusters at the bottom of the dendrogram show the most alignment, and the height between clusters shows the relative differences. Horizontal distances do not convey meaning.

The dendrogram indicates three paradigmic families of assessments. The most closely aligned assessments in terms of how they assessed cognitive operations and cognitive demand were (E) and (H), then (G) and then (F). (B) and (D) were aligned, and separately (A) and (C) were aligned.

Whether a position in paradigmic family groupings is seen as good or bad depends of course on the characteristics of the associated family members. The constituent factors of cognitive operations and cognitive demand are key to good assessments. (B) and (A) had relatively good proportions of cognitive operations, also cognitive demand, and therefore those assessments serve as positive comparative assessment to be aligned with. Therefore similarities have to be based on whether it is considered good or bad to have the constituent profiles of the chosen features.

Table 27: Cognitive operation v question type

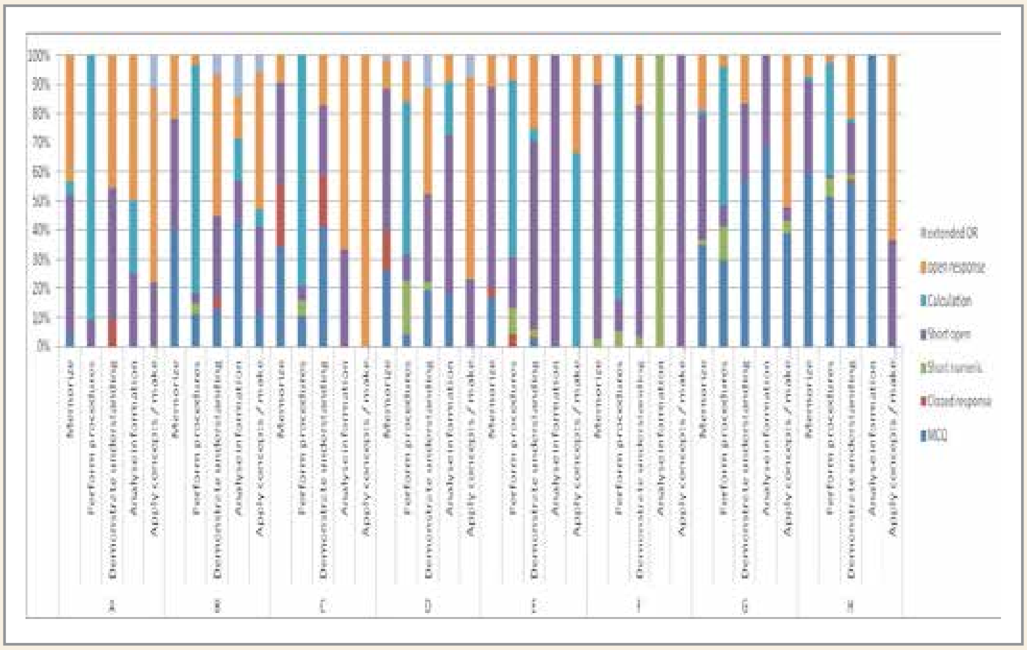
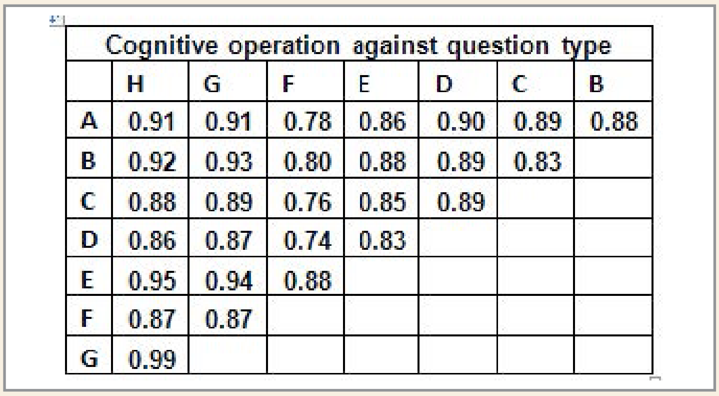


Table 28: Cognitive operation against question type



Cognitive operation against question type

	H	G	F	E	D	C	B
A	0.91	0.91	0.78	0.86	0.90	0.89	0.88
B	0.92	0.93	0.80	0.88	0.89	0.83	
C	0.88	0.89	0.76	0.85	0.89		
D	0.86	0.87	0.74	0.83			
E	0.95	0.94	0.88				
F	0.87	0.87					
G	0.99						

Figure 10: Cognitive operation against question type

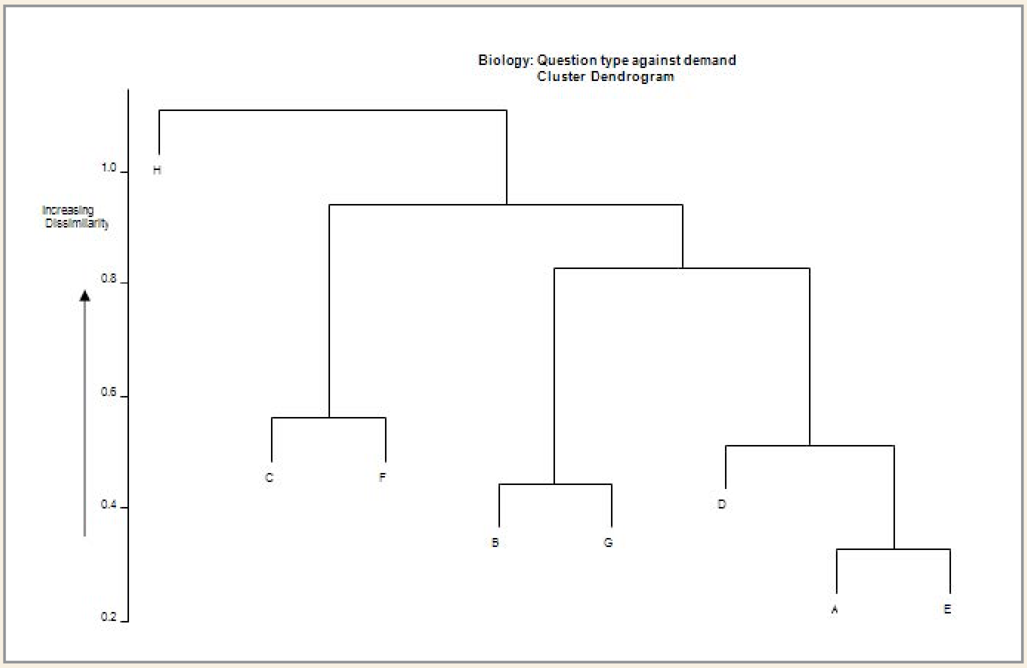


Table 29: Question type v cognitive demand

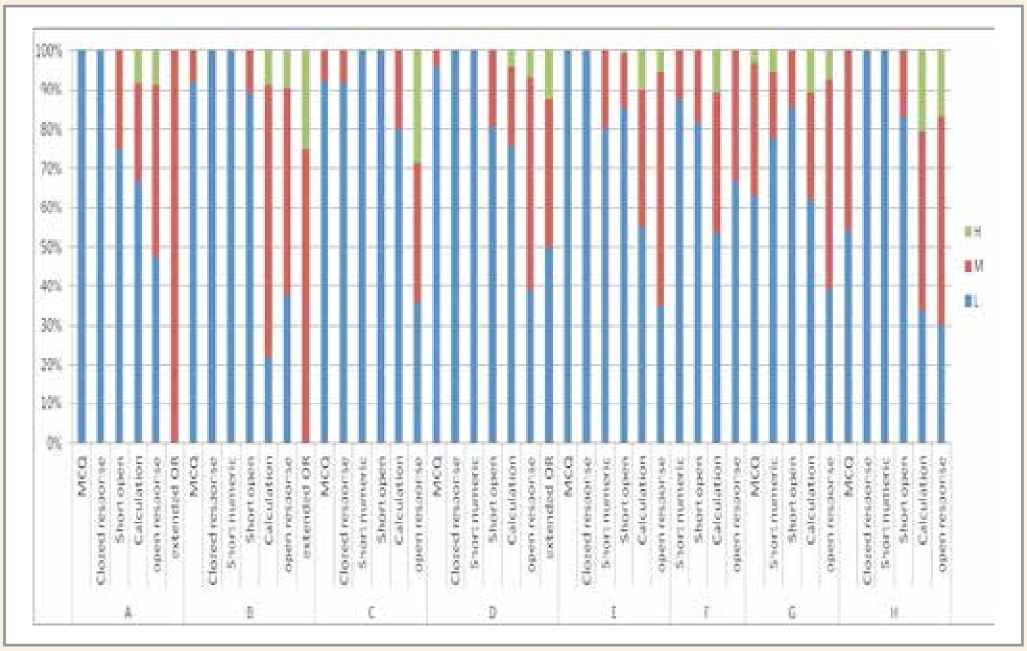
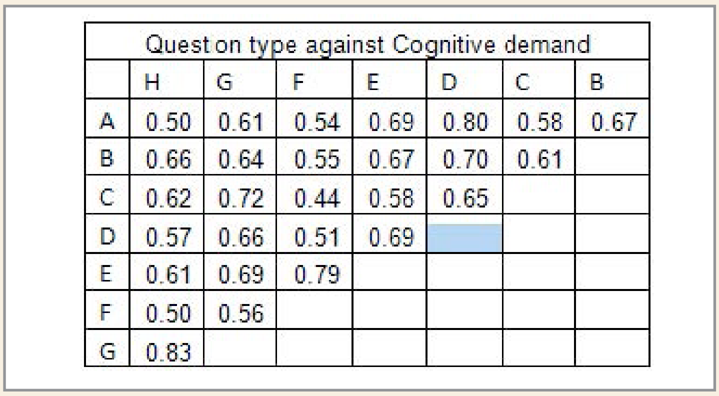
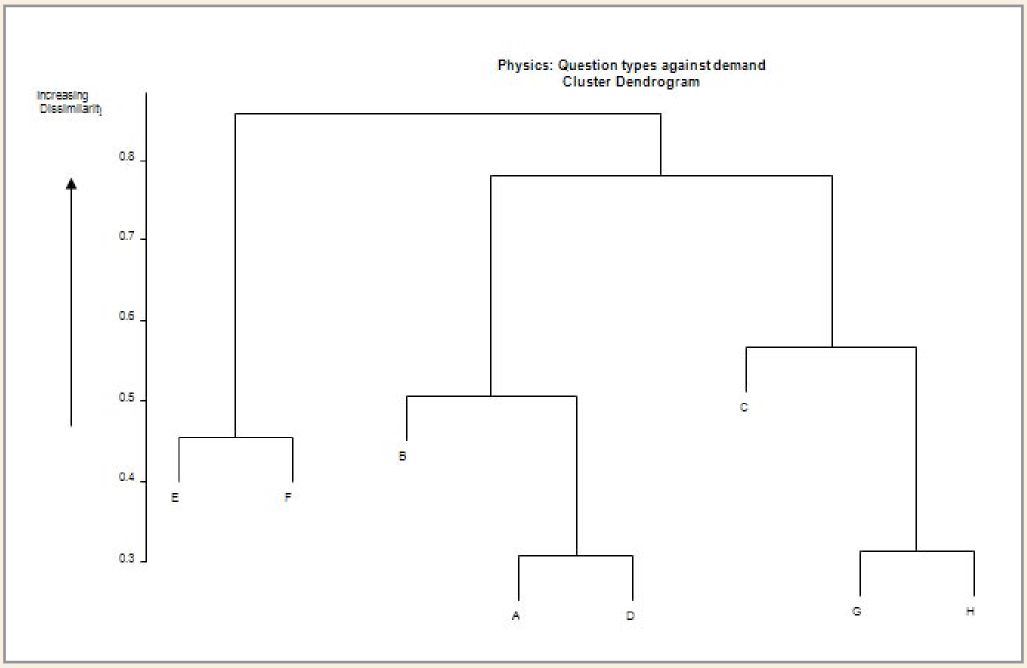


Table 30: Question type against cognitive demand



Quest on type against Cognitive demand							
	H	G	F	E	D	C	B
A	0.50	0.61	0.54	0.69	0.80	0.58	0.67
B	0.66	0.64	0.55	0.67	0.70	0.61	
C	0.62	0.72	0.44	0.58	0.65		
D	0.57	0.66	0.51	0.69			
E	0.61	0.69	0.79				
F	0.50	0.56					
G	0.83						

Figure 11: Question types against demand dendrogram



Tables 27 and 29 show profiles of how types of assessment tasks or processes students are asked to perform and their cognitive demand are linked to the question types used. These tables show that a broad range of question types was used to assess various cognitive operations. Open and more extended question types were more likely, but no guarantee to assess higher-order thinking skills. The dendrograms (Figures 10 and 11) show different clustering arrangements on the basis of both of these conditions.

Key findings for Physics

The key comparative features of Physics assessments across the selected examination boards and jurisdictions were:

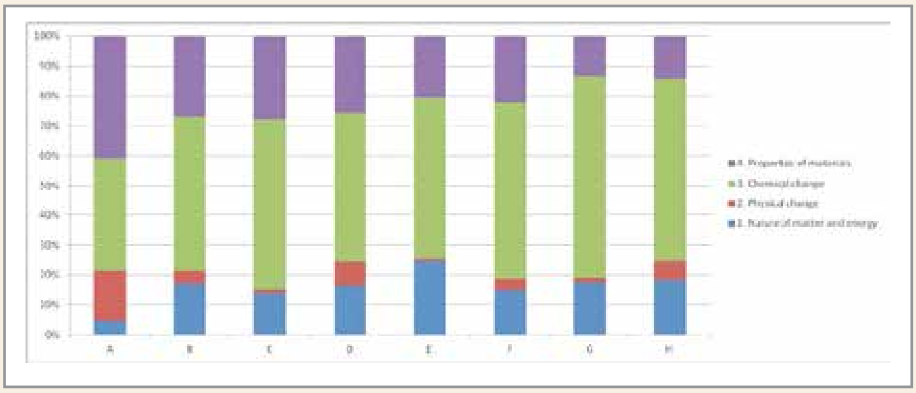
- In general, the assessment of physics was fairly well distributed between four of the content standards – 'Forces and motion', 'Light, sound and waves', 'Electricity and magnetism' and 'Energy and matter'.
- The only content standard area that was not assessed consistently was 'Earth and space'. (A), (E) and (G) did not assess this area at all and only (C) assessed it in even proportion to the other content areas at 21%.
- Sampling strategies, if used, were significantly different across jurisdictions in terms of the emphasis of physics content areas.
- The cognitive operation of 'memorise' ranged from 19% (G) to 39% (A).
- 'Performing procedures', which is largely the assessment of scientific enquiry skills, ranged from 19% (A) to 41% (H).
- 'Demonstrating understanding' ranged from 19% (A) to 35% (G).
- Higher-order cognitive operations ('analyse information and apply concepts/make connections) were in general assessed very little, e.g. 2% (E) and 3% (F). They were far better represented in A (22%), B (21%) and D (21%).
- Low cognitive-demand items dominate the assessments, ranging from 53% (H) to 82% (C). The proportions of medium-demand items ranged from 25% (C) to 41% (H).
- In general, there were few items at high demand across all jurisdictions, ranging from 2 to 6%.

The alignment indices, using Porter's Alignment Index equation, were applied to aspects of assessments themselves (rather than the way Porter used them – between content standards and assessment) as the construction and purposes of content standards are not consistent internationally. They show how closely all the jurisdictions are related to each other in terms of the three correlated analyses, and the accompanying dendrograms visually illustrate these relationships in terms of clusters or families of similar features. It can be seen that on the basis of this international comparative assessment research, there were significant similarities and differences between groups of assessments. In conjunction with the bar chart analyses, it has been possible to explore at item level the underlying reasons for similarities and differences between assessments, and be informed by good practice.

On the basis of reasonably good proportions of cognitive operations and cognitive demand, (A) and (B) serve as positive comparative assessments to be aligned with, with the caveat that there could and should be far better coverage, if designed for.

Chemistry international comparisons

Table 31: Content standards



It can be seen from Table 31 that there was considerable variation in the proportions of questions from the four content areas across national and international assessments. Linear assessments usually use sampling strategies over time to ensure content representation. As this study was purposed to look across a number of examination boards and jurisdictions over one or two examination series, the content coverage may not be generalisable. However, on the basis of this evidence there were significant differences in the emphases of different aspects of chemistry assessed in examinations.

The assessment content standard that dominated the selected chemistry assessments was 'chemical change' coverage across countries varying from approximately 40% (A) to 60% (H). Assessment items on 'Matter and energy' and 'Properties of materials' were relatively well represented with a few exceptions. It can also be seen that the chemistry area of 'Physical change' was barely assessed at all, ranging from 1% (C) to 18% (A). Whether these differences are by design or a result of some form of random sampling would need to be established before any conclusions could be made about under or over representation. How differing examination boards or jurisdictions create or maintain a sampling strategy with an associated justification and rationale is a key assessment issue in terms of validity, reliability, standard setting and maintenance.

Table 32: Cognitive operation

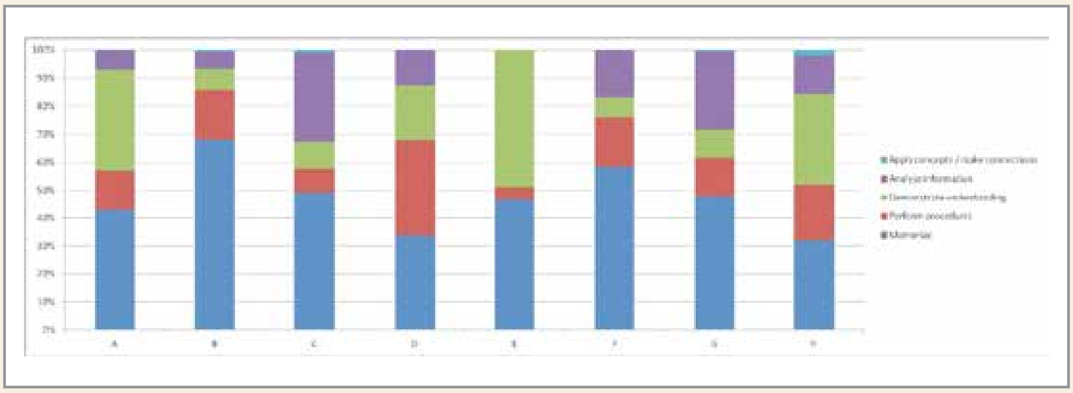


Table 32 shows the proportion of cognitive operations assessed in the selected Chemistry examinations. This mapping category was a key indicator of the design and profile of Chemistry assessments. The five Chemistry cognitive operations are descriptions of the types of tasks or processes that chemistry examiners and practitioners have agreed should be evident in assessments (more expanded detail on the types of chemistry procedures subsumed into each cognitive operation can be found on pages 4–5 – these are generic across all the sciences). The striking finding was the high amount of variation across all the cognitive operations. 'Memorise'-based items dominated coverage, ranging from 33% (D) to 68% (B). 'Demonstrate understanding' was variably covered, ranging from 5% (F) to 49% (E). Apart from (E) at 3%, 'performing procedures', focusing on enquiry-based skills were well represented in the assessments, particularly in (D) at 34%.

Those cognitive operations that should employ higher-order thinking skills ('analysing information' and 'applying concepts and making connections') had quite different profiles. Only (E) did not assess either of them at all. 'Analysing information' was far better represented out of the two, reaching 30% in (C). Applying concepts was represented very little, 1–2% in (H), (E) and (B), and in all other jurisdictions not at all.

Table 33: Cognitive demand

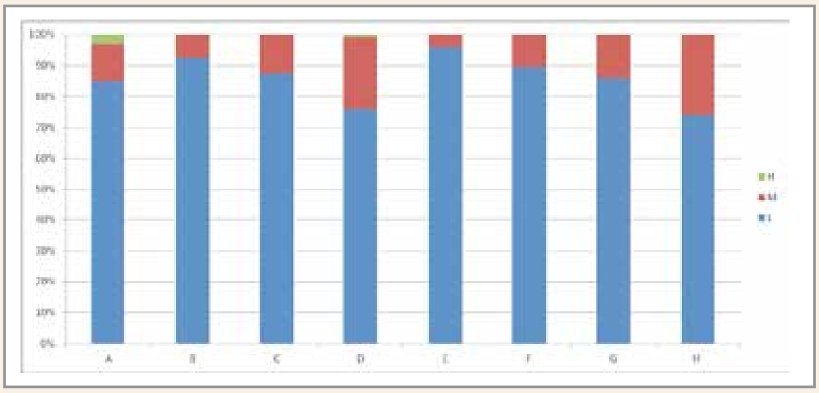


Table 33 shows the profile of assessment items of the basis of cognitive demand. More detail on how the cognitive demand criteria were established and operationalised is given on pages 5–8, however in short, this category describes the interrelation between the complexity of assessment items, and the expectations of their associated mark schemes and content standards. In Chemistry, low-demand items dominated the assessments, with a range between 73% (H) and 96% (C). There were three main reasons for this: mostly the items involved straightforward one-step procedures, or the mark schemes were constructed atomistically so that there was no requirement for any linked steps to gain credit **or** the content standards were so specifically described that items required no application, just repetition of a routine procedure. Items of a medium cognitive demand were therefore fairly low in proportion, ranging from 4% (E) to 27% (H). Highly cognitively demanding items were a rarity, hardly registering at all.

Table 34: Queston type

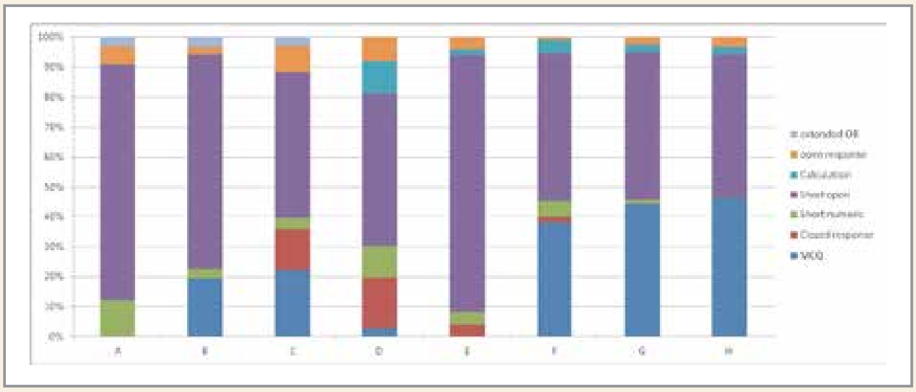


Table 34 shows how different examination boards and jurisdictions made use of different question types in chemistry. It can be seen that there was considerable variance. Some international areas made significant use of multiple-choice questions (MCQs), for example (H) and (G), others made no use of these question types, e.g. (A). The table shows that short open questions dominated assessments, ranging from 48% (H) to 80% (A). Surprisingly, numeric and calculations questions were relatively few and far between. The under-representation of mathematics in science assessments was an issue, and while the content representation of these assessments may explain some of the absence of calculations, it was an identifiable gap in a number of jurisdictions.

Table 35: Cognitive operation v cognitive demand

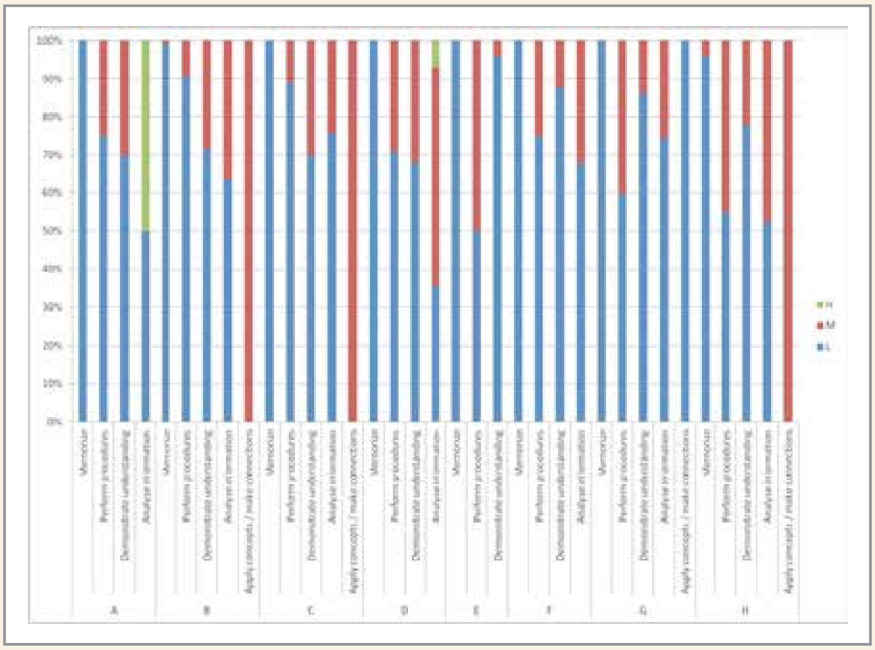
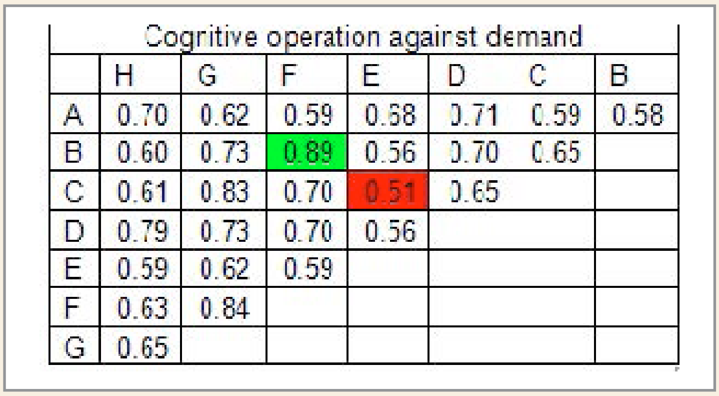


Table 36: Cognitive operation against demand



Cognitive operation against demand

	H	G	F	E	D	C	B
A	0.70	0.62	0.59	0.68	0.71	0.59	0.58
B	0.60	0.73	0.89	0.56	0.70	0.65	
C	0.61	0.83	0.70	0.51	0.65		
D	0.79	0.73	0.70	0.56			
E	0.59	0.62	0.59				
F	0.63	0.84					
G	0.65						

Figure 12: Cognitive operation against demand dendrogram

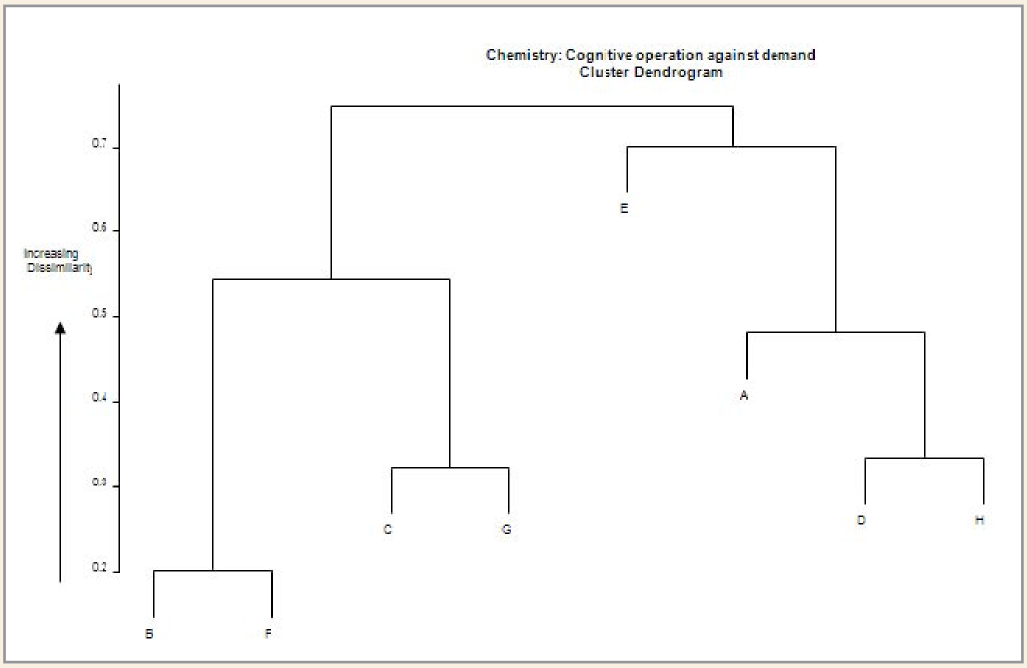


Table 35 shows the correlation between the cognitive operations (the types of tasks or processes) and the cognitive demand of chemistry assessments. It might be assumed that chemistry assessment items would become cognitively more demanding as cognitive operations move from those requiring memory skills across to those involving solving non-routine problems and making connections.

Table 35 demonstrates this to be a simplistic assumption. Although memory items were almost entirely of low demand, Table 35 also indicates that cognitive demand cannot be assumed by the type of assessment task. The almost complete lack of high-demand items reinforces this point. Most examination boards and jurisdictions presented a mixed profile of types of tasks to their low and medium cognitive demand.

The chemistry dendrogram (Figure 12) visually represents the tables of alignment indices. It shows the similarities between the relationship between cognitive operations and cognitive demands across jurisdictions. Close similarities are shown as clusters. The clusters at the bottom of the dendrogram show the most alignment, and the height between clusters shows the relative differences. Horizontal distances do not convey meaning.

There appear to be two paradigmic families of assessments, separated into the clusters shown above. The most closely aligned jurisdictions were (B) and (F), followed by (C) and (G). (D) and (H) were aligned, followed by (A), and weakly by (E).

Whether a position in paradigmic family groupings is seen as good or bad depends of course on the characteristics of the associated family members. The constituent factors of cognitive operations and cognitive demand are key to good assessments. (H) probably had the best profile of cognitive operations and associated cognitive demand and therefore Hong Kong assessments serve as positive comparative assessments to be aligned with. Therefore similarities have to be based on whether it is considered good or bad to have the constituent profiles of the chosen features. In Chemistry therefore (D) had a positive relationship in comparison with (H), followed by (A). In general the profiles of cognitive operations and cognitive demand indicated that there is general room for improvement both at home and abroad.

Table 37: Cognitive operation v question type

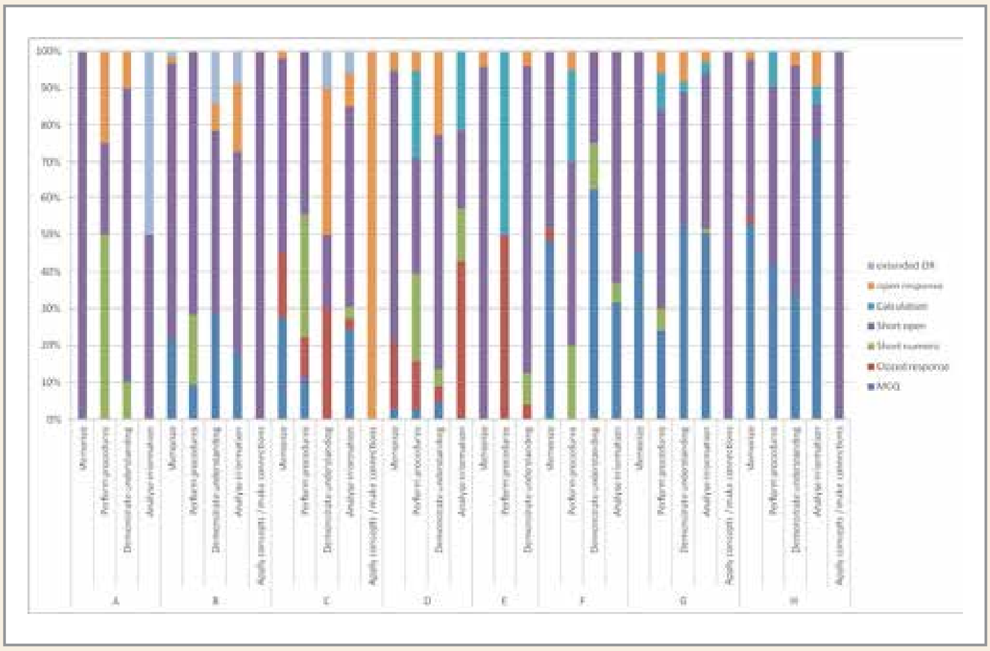
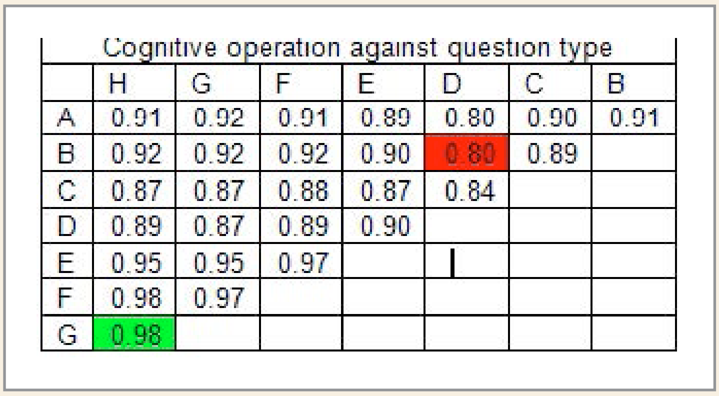


Table 38: Cognitive operation against question type



Cognitive operation against question type

	H	G	F	E	D	C	B
A	0.91	0.92	0.91	0.89	0.80	0.90	0.91
B	0.92	0.92	0.92	0.90	0.80	0.89	
C	0.87	0.87	0.88	0.87	0.84		
D	0.89	0.87	0.89	0.90			
E	0.95	0.95	0.97				
F	0.98	0.97					
G	0.98						

Figure 13: Cognitive operation against question type dendrogram

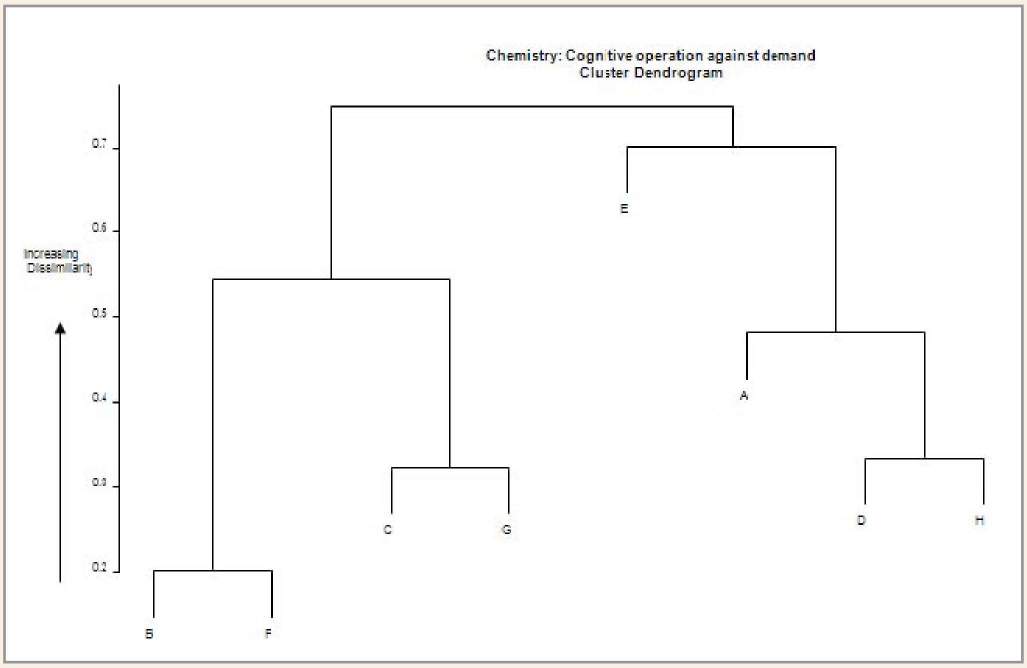


Table 39: Question Type v cognitive demand

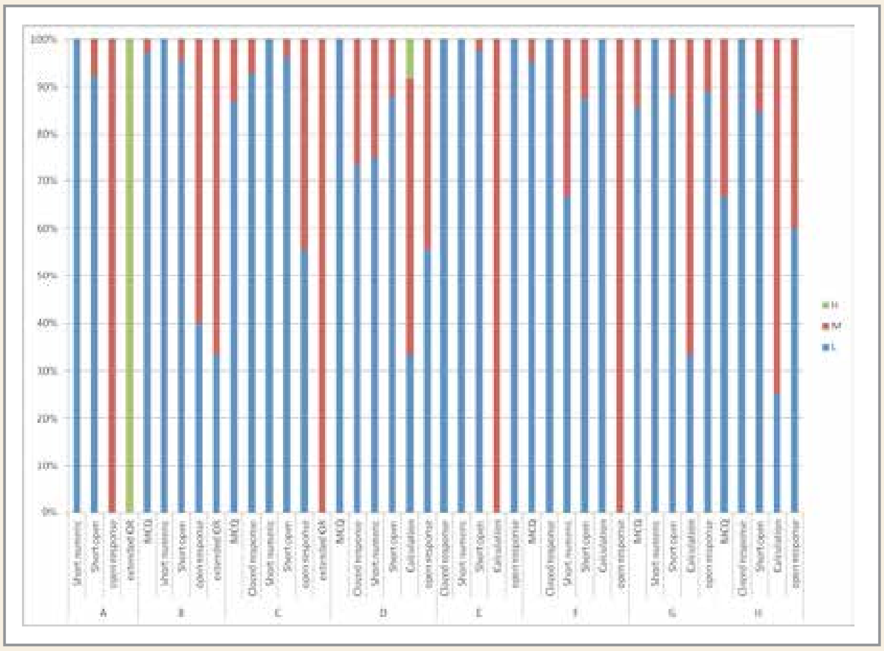
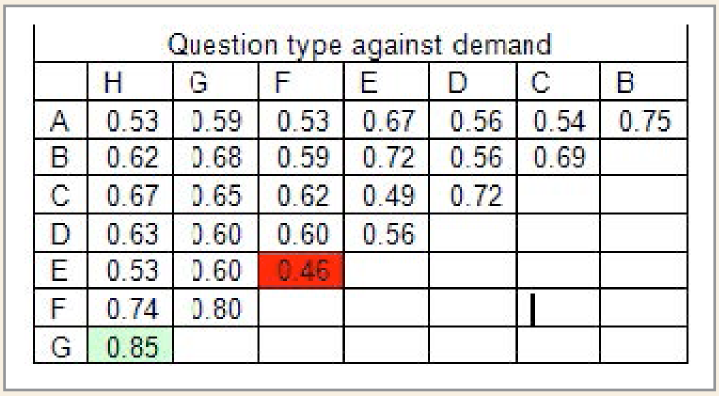


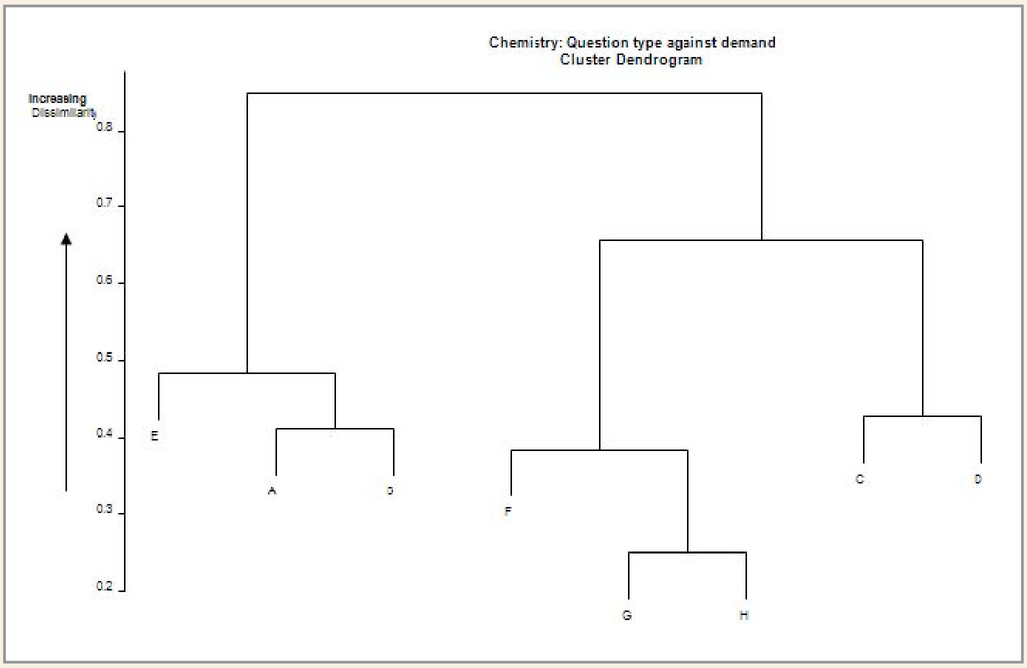
Table 40: Question type against demand



Question type against demand

	H	G	F	E	D	C	B
A	0.53	0.59	0.53	0.67	0.56	0.54	0.75
B	0.62	0.68	0.59	0.72	0.56	0.69	
C	0.67	0.65	0.62	0.49	0.72		
D	0.63	0.60	0.60	0.56			
E	0.53	0.60	0.46				
F	0.74	0.80					
G	0.85						

Figure 14: Question type against demand dendrogram



Tables 37 and 39 show profiles of how types of assessment tasks or processes students are asked to perform and their cognitive demand are linked to the question types used. These tables show that there was some range of question types used to assess various cognitive operations, however, the overall question types used in chemistry were more limited than biology and physics. The dendrograms (Figures 13 and 14) show different clustering arrangements on the basis of both of these conditions.

Key findings for Chemisty

The key comparative features of Chemistry assessments across the selected examination boards and jurisdictions were:

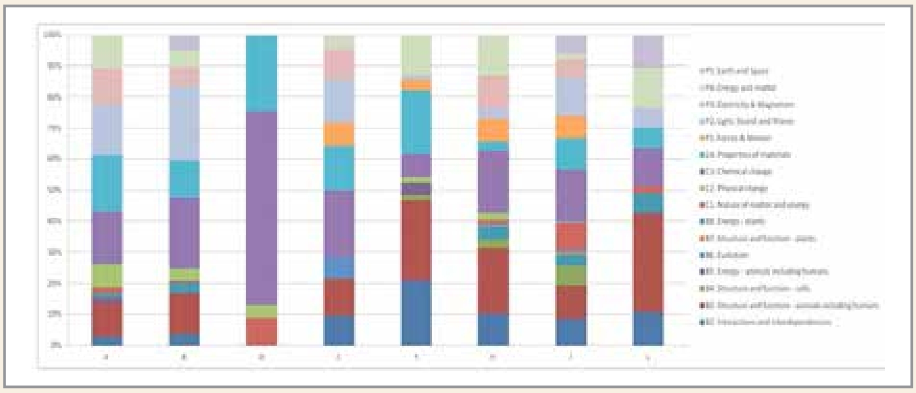
- In general, the assessment of Chemistry was dominated by the content standards covering 'chemical change' ranging from 40% (A) to 60% (H). 'Matter and energy' and 'Properties of materials' were relatively minor domains. 'Physical change' was minimally assessed, ranging from 1% (C) to 18% (A).
- Sampling strategies, if used, were significantly different across jurisdictions in terms of the emphasis of chemistry content areas.
- The cognitive operation of 'memorise' was dominant in the assessments, 'Memorise'-based items dominated coverage, ranging from 33% (D) to 68% (B).
- 'Demonstrate understanding' was variably covered, ranging from 5% (F) to 49% (E).
- Apart from (E) at 3%, 'performing procedures', focusing on scientific enquiry skills were well represented in the assessments, particularly in (D) at 34%. This coverage was significantly more represented in Chemistry than in Biology.
- Those cognitive operations that should employ higher-order thinking skills (analysing information and applying concepts and making connections) had quite different profiles. Only (E) did not assess either of them at all. Analysing information was far better represented out of the two, reaching 30% in (C). Applying concepts was represented very little, 1–2% in (H), (E) and (B), and in all other jurisdictions not at all.
- In Chemistry low-demand items dominated the assessments, with a range between 73% (H) and 96% (C).
- Items of a medium cognitive demand were fairly low in proportion, ranging from 4% (E) to 27% (H). Highly cognitively demanding items were a rarity, hardly registering at all.

The alignment indices, using Porter's Alignment Index equation were applied to aspects of assessments themselves (rather than the way Porter used them – between content standards and assessment) as the construction and purposes of content standards are not consistent internationally. They show how closely all the jurisdictions are related to each other in terms of the three correlated analyses, and the accompanying dendrograms visually illustrate these relationships in terms of clusters or families of similar features. It can be seen that on the basis of this international comparative assessment research, there were significant similarities and differences between groups of assessments. In conjunction with the bar chart analyses, it has been possible to explore at item level the underlying reasons for similarities and differences between assessments, and be informed by good practice.

On the basis of reasonably good proportions of cognitive operations and cognitive demand, (H) served as the most positive comparative assessment to be aligned with, with the caveat that there could and should be far better coverage of high cognitive demand items in Chemistry. Using this comparator, (D) compared positively, followed by (A).

Science international comparisons

Table 41: Content standards



It can be seen from Table 41 that there was a large number of content standards to map. This content standard list was constructed from the content strands of Biology, Chemistry and Physics, totalling 17 categories and therefore difficult to easily represent. A cursory view indicates significant variation of coverage across national and international assessments and in particular the varying proportions of Biology, Chemistry and Physics content areas. There are however some similarities between representation emphases in science and the individual science subjects. For example, the structure and function of animals is the dominant biology area ranging from 9% (D) to 31% (H). Chemical change dominates chemistry coverage, ranging from 12% (H) to 61% (D). Physics coverage was the most erratic, with 'Light, sound and waves' and 'Energy and matter' being the major domains. Linear assessments usually use sampling strategies over time to ensure content representation. As this study was purposed to look across a number of examination boards and jurisdictions over one or two examination series, the content coverage evidenced may not be generalisable. However, on the basis of this evidence there are significant differences in the emphases of different aspects of science assessed in examinations.

Table 42: Cognitive operations

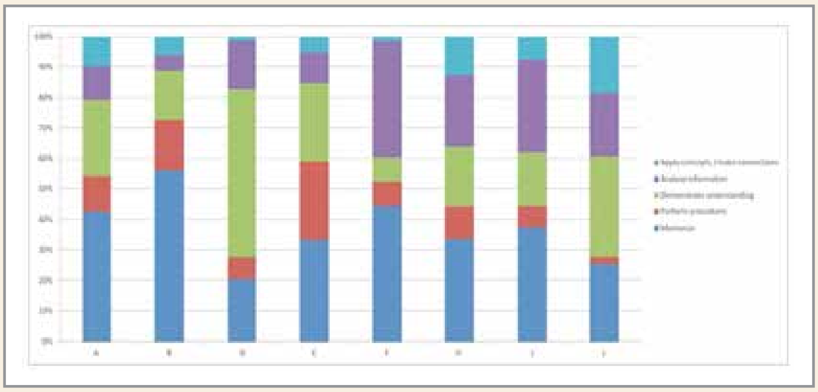


Table 42 shows the proportion of cognitive operations assessed in the selected science examinations. The five science cognitive operations are descriptions of the types of tasks or processes that science examiners and practitioners have agreed should be evident in assessments (more expanded detail on the types of science procedures subsumed into each cognitive operation can be found on pages 4–5 – they were generic across all the sciences).

The striking finding was the high amount of variance across all the cognitive operations. 'Memorise'-based items dominate coverage, ranging from 20% (D) to 56% (B). 'Demonstrate understanding' varied considerably, ranging from 8% (F) to 55% (D). Apart from (L) at 2%, 'performing procedures', focusing on enquiry-based skills were fairly well represented in the assessments, particularly in (E) at 25%.

Those cognitive operations that should employ higher-order thinking skills (particularly applying concepts and making connections' were represented more in science assessments than in the separate science subjects. Analysing information had better coverage, ranging from 5% (B) to 31% (J); the coverage of Applying concepts and making connections ranged from 2% (D and F) to 18% (L).

Table 43: Cognitive demand

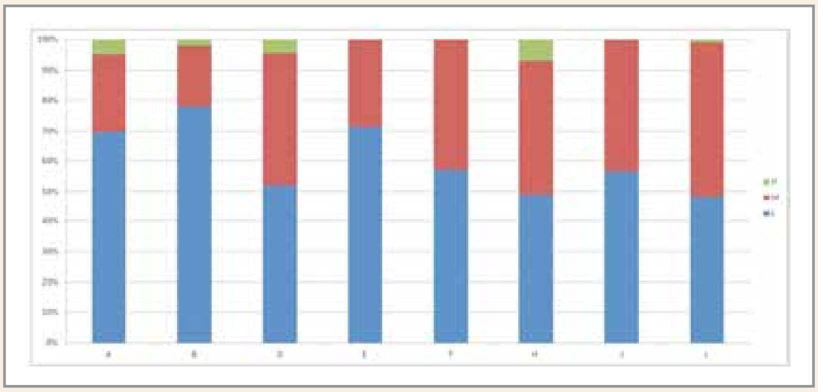


Table 43 shows the profile of assessment items of the basis on cognitive demand. More detail on how the cognitive demand criteria were established and operationalised is given on pages 5–8, however in short, this category describes the interrelation between the complexity of assessment items, and the expectations of their associated mark schemes and content standards.

In science, low-demand items dominated the assessments, with a range between 48% (L) and 78% (B). There are three main reasons for this; mostly the items involved straightforward one-step procedures, or the mark schemes were constructed atomistically so that there was no requirement for any linked steps to gain credit **or** the content standards were so specifically described that items required no application, just repetition of a routine procedure.

There was generally a good coverage of medium cognitive demand items, ranging from 29% (E) to 51% (L). Highly cognitively demanding items were a rarity, ranging from 0% (E, F and J) to only 3% (H).

Table 44: Question type

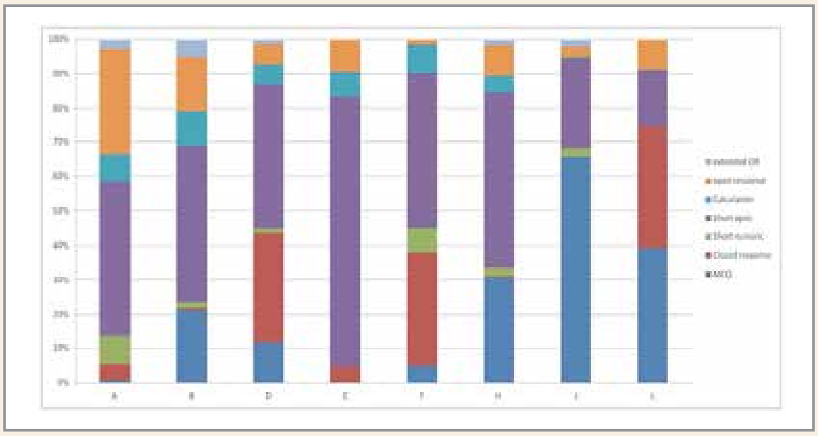


Table 44 shows how different examination boards and jurisdictions made use of different question types in science. It can be seen that there was considerable variance. Some international areas make significant use of multiple-choice questions (MCQs), particularly (J) at 66%, while others made no use of these question types. Apart from the use of MCQs, the table shows that short open questions dominate assessments, ranging from 18% (L) to 80% (E) while numeric and calculations questions were relatively few and far between. The under-representation of mathematics in science assessments is an issue, and while the content representation of these assessments may explain some of the absence of calculations, it is an identifiable gap in a number of jurisdictions.

Table 45: Cognitive operation v cognitive demand

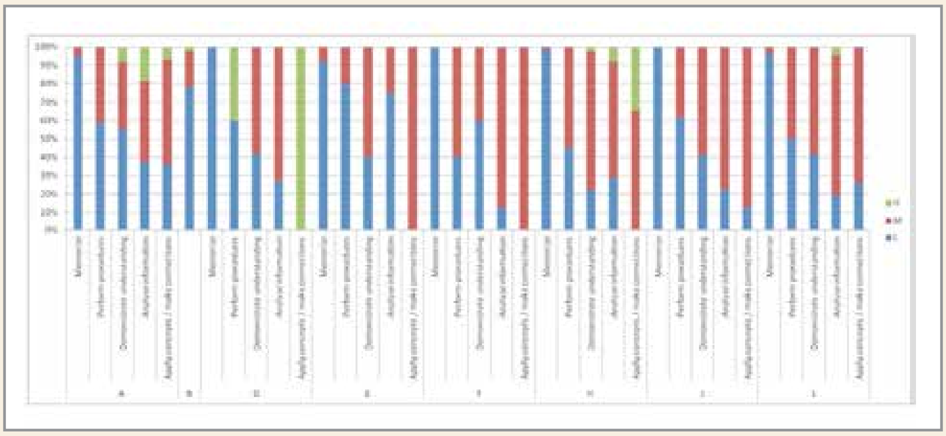
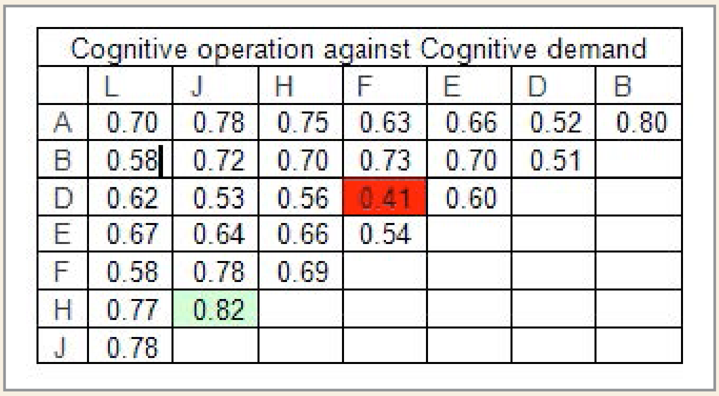


Table 46: Cognitive operation against cognitive demand



Cognitive operation against Cognitive demand

	L	J	H	F	E	D	B
A	0.70	0.78	0.75	0.63	0.66	0.52	0.80
B	0.58	0.72	0.70	0.73	0.70	0.51	
D	0.62	0.53	0.56	0.41	0.60		
E	0.67	0.64	0.66	0.54			
F	0.58	0.78	0.69				
H	0.77	0.82					
J	0.78						

Figure 15: Cognitive operation against demand dendrogram

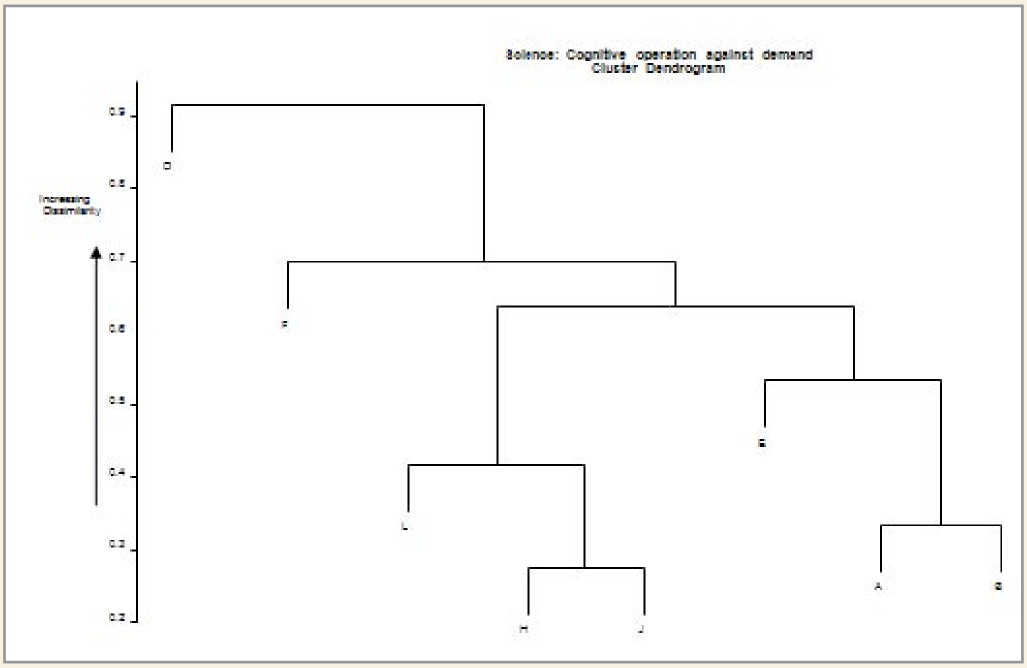


Table 45 shows the correlation between the cognitive operations (the types of tasks or processes) and the cognitive demand of science assessments. It might be assumed that science assessment items would become cognitively more demanding as cognitive operations move from those requiring memory skills across to those involving analysing and making connections.

Table 45 again indicates that this is not true and cognitive demand cannot be assumed by the type of assessment task. The almost complete lack of high-demand items reinforces this point. Most examination boards and jurisdictions presented a mixed profile of types of tasks to their low and medium cognitive demand.

The science dendrogram (Figure 15) visually represents the tables of alignment indices. It shows the similarities between the relationship between cognitive operations and cognitive demands across jurisdictions. Close similarities are shown as clusters. The clusters at the bottom of the dendrogram show the most alignment, and the height between clusters show the relative differences. Horizontal distances do not convey meaning.

Of all the dendrograms for mathematics and the sciences, the subject of science showed the most branched differences across jurisdictions. The most closely aligned jurisdictions were (H) and (J) followed by (L). Separately (A) and (B) showed close alignment, followed by (E). (D) and (F) were both independent outliers.

Whether a position in paradigmic family groupings is seen as good or bad depends of course on the characteristics of the associated family members. The constituent factors of cognitive operations and cognitive demand are key to good assessments. (H) probably had the best profile of cognitive operations and associated cognitive demand and therefore their assessments serve as positive comparative assessments to be aligned with. Therefore similarities have to be based on whether it is considered good or bad to have the constituent profiles of the chosen features. In science, therefore, (J) has a positive relationship in comparison with (H) and (L). In general, the profiles of cognitive operations and particularly cognitive demand indicate that there is a general room for improvement both at home and abroad.

Table 47: Cognitive operation v question type

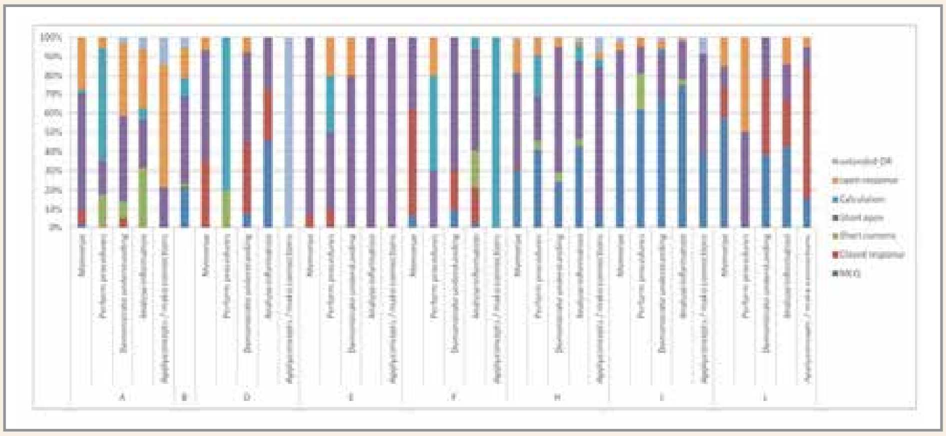
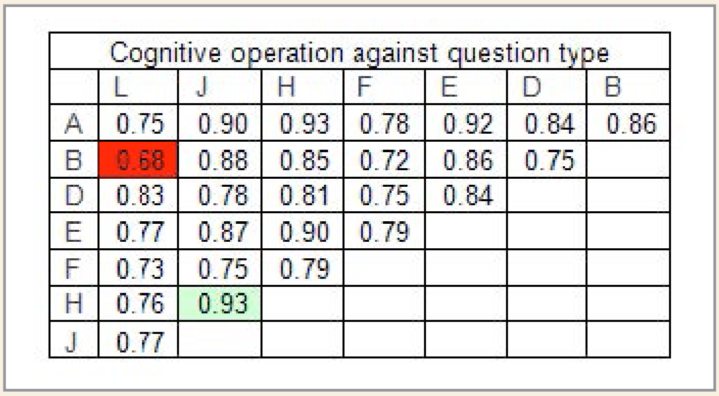


Table 48: Cognitive operation against question type



Cognitive operation against question type							
	L	J	H	F	E	D	B
A	0.75	0.90	0.93	0.78	0.92	0.84	0.86
B	0.68	0.88	0.85	0.72	0.86	0.75	
D	0.83	0.78	0.81	0.75	0.84		
E	0.77	0.87	0.90	0.79			
F	0.73	0.75	0.79				
H	0.76	0.93					
J	0.77						

Figure 16: Cognitive operation against question type

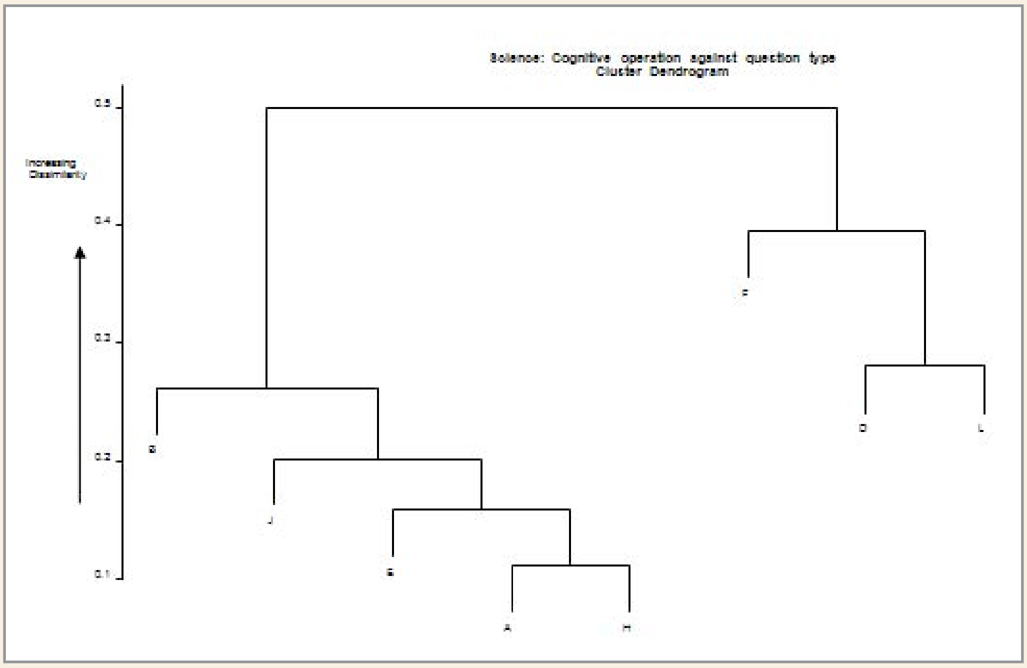


Table 49: Question type v cognitive demand

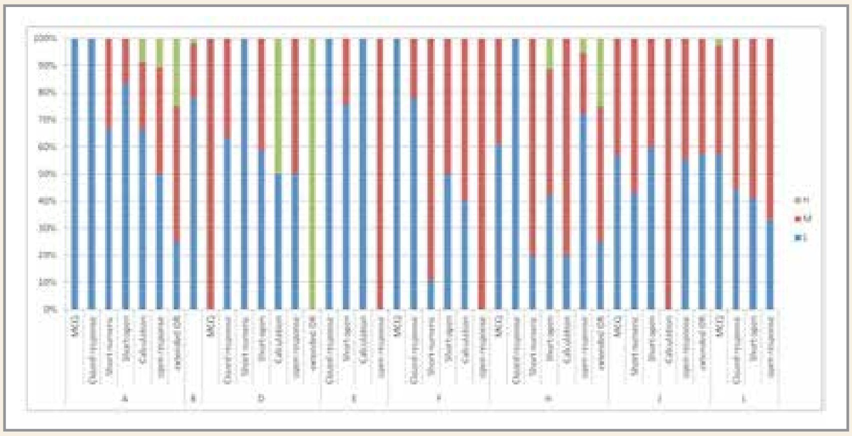
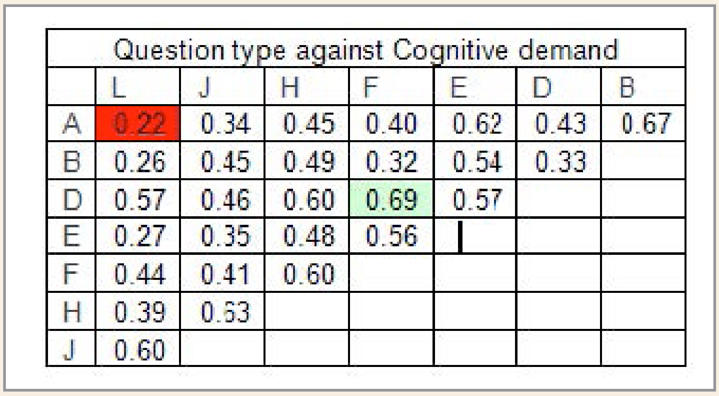
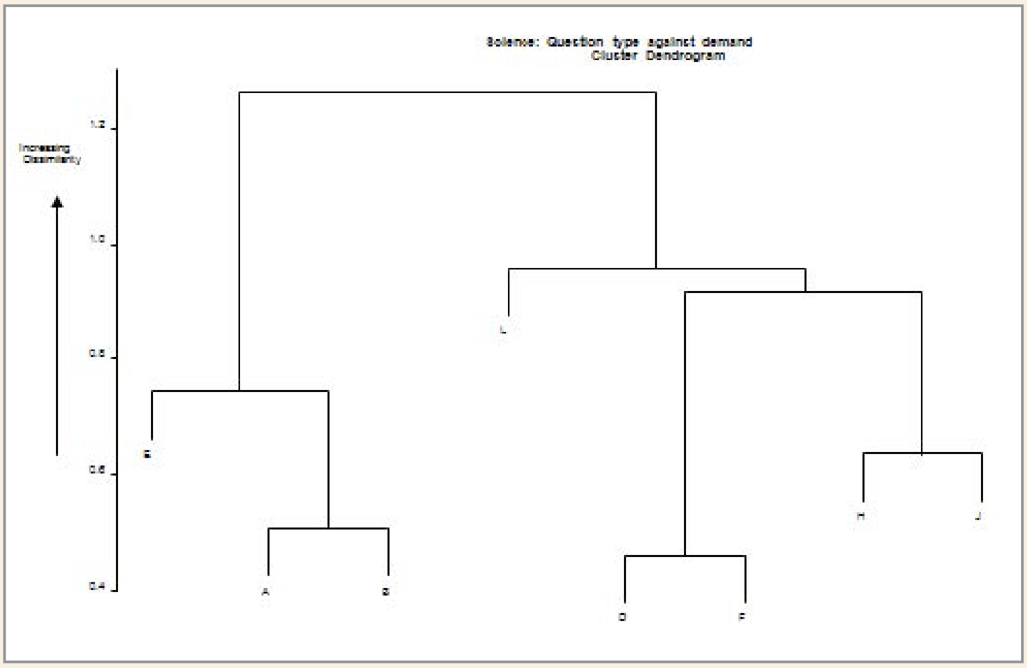


Table 50: Question type against cognitive demand



Question type against Cognitive demand							
	L	J	H	F	E	D	B
A	0.22	0.34	0.45	0.40	0.62	0.43	0.67
B	0.26	0.45	0.49	0.32	0.54	0.33	
D	0.57	0.46	0.60	0.69	0.57		
E	0.27	0.35	0.48	0.56			
F	0.44	0.41	0.60				
H	0.39	0.53					
J	0.60						

Figure 17: Question type against demand dendrogram



Tables 47 and 49 show profiles of how types of assessment tasks or processes students are asked to perform and their cognitive demand are linked to the question types used. These tables show that there was some range of question types used to assess various cognitive operations, however, the overall question types used in Science was more limited than Biology and Physics, and as described before, there weren't any high-demand items to compare against.

The dendrograms show different clustering arrangements on the basis of both of these conditions.

Key findings for Science

The key comparative features of Science assessments across the selected examination boards and jurisdictions were:

- As there are so many potential content standards in Science – as it subsumes Biology, Chemistry and Physics, it is difficult to generalise about comparative content coverage. However, the structure and function of animals was the dominant biology area ranging from 9% (D) to 31% (H). Chemical change dominated chemistry coverage, ranging from 12% (H) to 61% (D). Physics coverage was the most erratic, with 'Light, sound and waves' and 'Energy and matter' being the major domains.
- Sampling strategies, if used, were significantly different across jurisdictions in terms of the emphasis of science content areas.
- The striking finding was the high amount of variance across all the cognitive operations. 'Memorise'-based items dominate coverage, ranging from 20% (D) to 56% (B).
- 'Demonstrate understanding' also varied considerably, ranging from 8% (F) to 55% (D).
- Apart from (L) at 2%, 'Performing procedures', focusing on scientific enquiry skills were fairly well represented in the assessments, particularly in (E) at 25%.
- Those cognitive operations that should employ higher-order thinking skills (particularly applying concepts and making connections) were represented more in Science assessments than in the separate science subjects. 'Analysing information' had better coverage, ranging from 5% (B) to 31% (J); the coverage of 'Applying concepts and making connections' ranged from 2% (D and F) to 18% (L).
- In general, the representation of all the cognitive operations was better in science than in the separate sciences.
- In science, low-demand items dominated the assessments, with a range between 48% (L) to 78% (B).
- There was generally a good coverage of medium cognitive demand items, ranging from 29% (E) to 51% (L).
- Highly cognitively demanding items were a rarity, ranging from 0% (E, F and J) to only 3% (H).

The alignment indices, using Porter's Alignment Index equation, applied to aspects of assessments themselves (rather than the way Porter used them – between content standards and assessment) as the construction and purposes of content standards are not consistent. They show how closely all the jurisdictions are related to each other in terms of the three correlated analyses, and the accompanying dendrograms visually illustrate these relationships in terms of clusters or families of similar features. It can be seen that on the basis of this international comparative assessment research, there are significant similarities and differences between groups of assessments.

On the basis of reasonably good proportions of cognitive operations and cognitive demand, (H) served as the most positive comparative assessments to be aligned with in Science, with the caveat that there could and should be far better coverage if designed for, particularly in items of higher cognitive demand. (J) was therefore the most positively aligned jurisdiction to (H), and then (L).

It was of interest to note that the cognitive operations that are often cited as those to develop higher-order and transferable thinking skills were significantly better represented in Science than in the separate subjects. It is also worth pointing out that there were more available science assessments for 16-year-olds than for the separate sciences and therefore, was better represented in terms of international comparisons.

General discussion points

1. The five mapped categories of uniform content standards, cognitive operations, cognitive demand, question types and the number of marks awarded under these classifications allowed for a number of rich analyses to be carried out to show relative proportions of and between these conditions. Alignment indices and their accompanying dendrograms illustrated the relative similarities and differences between the jurisdictions, and indicated preferred or desired alignments.

2. Across Mathematics and the sciences, there were significant variations in uniform content standard coverage both in terms of the number and weighting of content standards assessed and their assessed proportions. Without knowing what sampling strategies were in place in the selected jurisdictions, it would be unfair to make assumptions about content representation. However, appropriate content representation is a key assessment issue in terms of validity, reliability and the setting and maintenance of standards, and therefore it would be expected that all jurisdictions would sample content standards effectively over time. This should be monitored longitudinally, and needs to be incorporated into the design principles for new GCSEs.

3. A criticism of some Mathematics and Science curricula is the ‘mile wide and inch deep’ stigma. This inference has often been levelled as a reason that certain jurisdictions perform poorly in the content assessments of TIMSS. This research study suggests that this criticism was not well founded in the assessment evidence. The jurisdictions that had significantly less content representation in terms of mathematical or science areas did not show evidence of assessing a wider range of cognitive operations or at higher cognitive demands.

4. The mapping of the assessment of cognitive operations illustrated the predominance of memory and recall items across jurisdictions. However, there was significant variation between jurisdictions and subject areas; some jurisdictions seemed to employ more effective assessment design principles in terms of the range of cognitive operations used.

5. Those cognitive operations usually associated with higher-order thinking skills were not well represented in assessments. This was particularly evident for problem-solving skills in Mathematics and for applying concepts and making connections in Science. If desired, these skills and assessment items need to be focused on, promoted in terms of content standards and then represented appropriately in assessments.

6. The mapping of cognitive demand of assessments illustrated the dearth of high-demand assessment items across all jurisdictions. There were a number of common factors contributing to this evidence, including the lack of question complexity, the use of atomistic mark schemes and the content standard specificity. The evaluation of cognitive demand also exposed the assumption that higher-order cognitive operations must be of a higher cognitive demand. These two factors can be aligned, but often they are not. This is not necessarily a bad thing, as long as test developers understand the difference and design their assessments accordingly.

7. Cognitive operations and cognitive demand often work independently of each other, and therefore need to be actively designed into assessments. Their separation, in terms of assessment design, allows for more varied assessments to be targeted across the range of student ability.

8. It should be acknowledged that there is no value judgement in constructing assessments with memory or recall items at low cognitive demand if that is the desired intention. It probably should be the case however that assessed 'knowledge' should consist of a range of cognitive operations at a range of cognitive demands. It is highly unlikely that this will occur by chance, especially over time. The understanding and application of clear and high quality assessment design principles are required to achieve this.

9. The subject of Science in general had a stronger assessment profile than the separate sciences in terms of the range of cognitive operations and cognitive demands.

10. There was significant variation across jurisdictions on the use of question types. The analyses showed that although more closed question types can assess a range of cognitive operations and cognitive demands, they are less effective than more open responses.

11. The number of calculation items found in science assessments were relatively low, especially in Biology and Chemistry. This supports the view that there is a general lack of mathematical skills in science assessments.

Overall, the key finding of this research study has been that has been that good assessments do not happen by accident: they require good assessment design principles followed by good practice in terms of assessment question and mark scheme construction. Having dissected national and international assessments in some detail, desired assessment principles and outputs can be developed and sustained using robust classification rubrics for all the required conditions of assessments: sampled content representation, a range of cognitive operations, identifiable cognitive demands and a range of question types purposed appropriately.

This comparative analysis did not reveal perfect assessments in Mathematics or in the sciences, however there were strong indications that certain jurisdictions were more designed and of higher quality than others in terms of content representation and the range of cognitive operations, cognitive demands and question types. Although an outlier in terms of their direct comparison to UK KS4 assessments, the assessments from (H) were generally indicated as the most positive assessments to be aligned with. These judgements were made on the basis of better distributions of cognitive operations and cognitive demands. Feedback from the raters re-enforced this empirical view.

This comparative analysis research study demonstrated how the four key conditions of content standards, cognitive operations, cognitive demand and question types can be designed to create high-quality, reliable and valid assessment instruments. The coarse-grain interrogation of the reported conditions in this research allows for finer-grain examination of the assessment items themselves, which can then inform test developers to learn from best practice. With care, assessments can be designed to reflect curriculum aims and the intended content standards.

The intention need not be to directly replicate or mimic other jurisdictions; assessments need to be designed to fulfil national and subject specific aims. However, there is clearly good practice evident in ensuring that students across the ability range are assessed in a comprehensive yet appropriate manner. In essence, this research has reported on and advocates good assessment by design.

References

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: the classification of educational goals Handbook 1, Cognitive domain.* London: Longman.

Biggs, J. and K. Collis (1982). *Evaluating the quality of learning: The SOLO taxonomy.* New York, Academic Press.

Brookhart, S.M. & Nitko, A.J. (2011) 'Strategies For Constructing Assessments of Higher Order Thinking Skills'. In G. Schraw & D.R. Robinson (Eds), *Assessment of Higher Order Thinking Skills* (pp.327-359). North Carolina: IAP.

Clesham, R. (2012a). Proceedings from the Invited Symposium on A level Reform: Oxford University Centre for Educational Assessment (OUCEA), Sept. 2012.

Clesham, R. (2012b). Proceedings from the Association for Educational Assessment Conference (AEA-E), Berlin, November 2012.

Davis, S.L. & Buckendahl, C.W. (2011). 'Incorporating Cognitive Demand in Credentialing Examinations.' In G. Schraw & D.R. Robinson (Eds), *Assessment of Higher Order Thinking Skills* (pp.327-359). North Carolina: IAP.

DfE (2012). 'Review of the National Curriculum in England: what can we learn from the English, mathematics and science curricula of high performing jurisdictions?' Research Report, DFE RR178, February 2012.

Edwards, J. & Dall'Alba, G. (1981). 'Development of a scale of cognitive demand for analysis of printed secondary science materials.' *Research in Science Education*, 11, 158–170.

Ford, M. J. & Wargo, B.M. (2011 (submitted)). 'Dialogic framing of scientific content for conceptual and epistemic understanding.'

Marzano, R. J. and J. S. Kendall (2007). *The new taxonomy of educational objectives*. Thousand Oaks, CA, Corwin Press.

Millar, R. & Osborne, J. F. (Eds.). (1998). 'Beyond 2000: Science Education for the Future.' London: King's College London.

OECD (2009). Take the Test: Sample Questions from OECD's PISA Assessments.

OECD (2012). Draft Scientific Framework for PISA 2015 Science (submitted by Pearson).

Pollitt, A. Ahmed, A. & Crisp, V. (2007). 'The demands on examination syllabuses and question papers.' In P. Newton, J-A Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds), *Techniques for monitoring the comparability of examination standards*. 166–206. London: Qualification and Curriculum Authority.

Porter, A.C. (2002). 'Measuring the Content of Instruction: Uses in Research and Practice. Educational Researcher,' Vol 31, No 7, pp 3–14.

Webb, N.L. (1997). 'Criteria for alignment of expectations and assessments in mathematics and science education.' Washington, DC, Council of Chief State School Officers and National Institute for Science Education Research Monograph.