

Predicting item difficulty of science National Curriculum tests: The case of Key Stage 2 assessments

Yasmine H. El Masri^{a*}, Steve Ferrara^b, Peter W. Foltz^c and Jo-Anne Baird^a

^aOxford University Centre for Educational Assessment (OUCEA), Oxford, UK

^bIndependent consultant

^cUniversity of Colorado, Boulder CO , USA

* *Corresponding author*

Keywords

item difficulty, item demands, Key Stage 2, science assessments, coding framework, *Coh-Metrix*

Abstract

Predicting item difficulty is highly important in education for both teachers and item writers. Despite identifying a large number of explanatory variables, predicting item difficulty remains a challenge in educational assessment with empirical attempts rarely exceeding 25% of variance explained.

This paper analyses 216 science items of key stage 2 tests which are national sampling assessments administered to eleven year olds in England. Potential predictors (topic, sub-topic, concept, question type, nature of stimulus, depth of knowledge and linguistic variables) were considered in the analysis. Coding frameworks employed in similar studies were adapted and employed by two coders to independently rate items. Linguistic demands were gauged using a computational linguistic facility. The stepwise regression models predicted 23% of the variance with extended constructed questions and photos being the main predictors of item difficulty.

While a substantial part of unexplained variance could be attributed to the unpredictable interaction of variables, we argue that progress in this area requires improvement in the theories and the methods employed. Future research needs to be centred on improving coding frameworks as well as developing systematic training protocols for coders. These technical advances would pave the way to improved task design and reduced development costs of assessments.

Introduction

Understanding what makes an assessment task more or less challenging for students is of prime importance in education on pedagogic and testing levels. On the pedagogic level, teachers should provide students with tasks that match their abilities and support their learning (Black & Wiliam, 1999). In other words, teachers need to be able to select particular questions with features that promote specific learning goals. They also need to anticipate problems generated by some aspects of the question that may for instance overly challenge students or confuse them. On a more technical level, item writers need to know what affects the level of difficulty of questions to manipulate test demands and best reflect the construct assessed (Pollitt, Ahmed & Crisp, 2007). They need to design questions that best elicit students' understanding and the skills they acquired (Ahmed & Pollitt, 2007). They also need to consciously exclude construct-irrelevant elements that do not pertain to the attribute measured (i.e. construct-irrelevant features) and that may weaken the validity of the question (Messick, 1993) such as ensuring that a science item does not impose high reading demands. Pre-testing of items is very expensive and is not always an option, especially in cases where item security is a concern.

In this paper, we use data of assessment questions that are pre-tested before administration to draw useful lessons about factors that affect their difficulty in an aim to build stronger theoretical models of task difficulty that could inform practice and test design. This way, teachers have a sound framework upon which to base classroom assessment development. In addition, test developers would discard fewer unsuitable items after live assessment.

A large number of studies examining what makes questions difficult have been carried out since the mid-eighties (Ahmed & Pollitt, 1999; Bramley et al., 1998; Ferrara & Duncan, 2011; Ferrara et al., 2007; Ferrara et al., 2011; Fisher-Hoch et al., 1997; Hambleton & Jirka, 2006; Pollitt et al., 2007; Pollitt et al., 1985; Pollitt & Ahmed, 2000). In Britain, Pollitt et al. (1985) analysed Scottish examinations. In the *Question Difficulty Project*, English high-stakes examinations in various subjects were investigated (Ahmed & Pollitt, 1999; Bramley et al., 1998; Fisher-Hoch et al., 1997; Pollitt et al., 1998). These studies identified a number of variables associated with the level of difficulty of the questions (e.g. type of question, technical terms, reading load, etc.). However, results were inconclusive as manipulation of items has not consistently been successful. For instance, Bramley et al. (1998) and Fisher-Hoch et al. (1997) found that altering science and mathematics items, respectively, with the

intention of making them easier sometimes led to harder items. It is worth noting that most of the aforementioned research is in the form of unpublished reports, lacking in detail in the description of methods, such that they do not provide sufficient information to evaluate the quality of results.

In recent studies in the USA, Ferrara and colleagues (Ferrara & Duncan, 2011; Ferrara et al., 2007, 2011) developed a framework of item response demands that classified item demands into two categories: cognitive and linguistic. The framework has been applied to state-wide school assessments in various subjects: e.g. mathematics (Ferrara et al., 2011) and science (Ferrara & Duncan, 2011). The application of the framework was effective in examining the progression of demands in science items across year levels as well as the alignment of item demands in science assessments with the content standards. Nevertheless, the authors admitted that more empirical data need to be collected to evaluate the usefulness of the framework proposed.

Despite huge efforts to identify explanatory variables, predicting item difficulty remains a challenge in educational assessment for both teachers and examiners (Hambleton & Jirka, 2006; Impara & Plake, 1998; Meyer, 2003; Wolf, Smith, & Birnbaum, 1995). It has been argued that poor predictability of item difficulty was largely due to the complex and unpredictable interaction of the various item features and demands (Bramley et al., 1998; Fisher-Hoch et al., 1997; Hambleton & Jirka, 2006; Jones, 1993; McLone & Patrick, 1990; Pollitt et al., 2007). The most successful attempt to predict item difficulty using a number of variables only managed to explain 20-23% of the variance in a range of subjects (Ferrara & Steedle, 2014). This leaves around 80% of variance unexplained. Thus, we acknowledge that research is still needed to improve methods of predicting item difficulty.

In this paper, we attempt to predict item difficulty of Key Stage 2 (KS2) science tests using variables outlined in the literature. KS2 tests are national science sampling assessments administered to 11 year olds in England and Wales. They have been less studied in comparison with other assessments such as the General Certificate of Secondary Education, GCSEs or A-levels. Some research explored validity issues related to KS2 assessments (Stobart, 2001) while other studies investigated reliability concerns related to these assessments (Maughan, Styles, Lin, & Kirkup, 2012; Newton, 2009). Shorrocks-Taylor and Hargreaves (1999) explored language issues associated with the level of difficulty of KS2

mathematics questions, yet no study analysed features affecting the level of difficulty of KS2 science items.

More recently, He et al. (2014) investigated measurement invariance of KS2 science tests in England as part of the *Assessment Validity Programme* run by the Office of Qualifications and Examinations Regulation (Ofqual). The authors noted that Rasch analyses of the data suggested that the 2010 science series, and more so the 2011 series, were too easy. In other words, the science assessments in 2010 and 2011 did not target the student population well with many items being too easy for the student sample. Normally, items that are too easy for a population of test-takers do not contribute much to the measurement of individual pupil performance; nevertheless, there is often a rationale behind including them in a test such as setting the overall difficulty of the assessment to a specific intended level (Maughan et al., 2012). Identifying variables which make some items very easy (while others more difficult) is crucial in the process of the level of difficulty of the entire test. In addition, He et al. (2014) investigated differential item functioning (DIF) in the science items of the 2010 and 2011 series; that is, they examined whether some items were systematically more difficult for specific subgroups (gender, ethnic, language and socio-economic subgroups). DIF techniques are often used to detect items which could result in bias across groups. He et al. (2014) identified DIF in several items with one item exhibiting particularly a large amount of DIF. The authors called for a further examination of the items to identify the source of DIF across groups in a hope to improve future item design.

The purpose of this paper is threefold: (1) understand the variables influencing the level of difficulty of KS2 items; (2) contribute to the literature on models of item difficulty and; (3) highlight the methodological challenges encountered in such analyses. In this paper, we argue that much of the challenge in predicting item difficulty is methodological. The paper is divided into four sections. First, we review previous literature identifying variables associated with item difficulty in science assessments. In the second section, we provide an overview of KS2 science sampling tests in England, describe the data and present the methodology adopted including data analysis procedures. In the third section, we provide results of the regression analyses as well as the amount of variance predicted based on the selected variables. In the last section, we discuss the results as well as the limitations of this study. We conclude with avenues for future research.

What makes science questions difficult?

What makes a science question difficult? An intuitive response would point to higher demands placed on examinees. In other words, more demanding questions tend to be more difficult. However, as is outlined later, this association is much more complex (Pollitt et al., 2007). Before delving into the discussion any further, it would be worth clarifying some semantic confusion around the concepts of item difficulty and item demand which are often used interchangeably in the education literature.

1. Distinguishing item difficulty and item demands

According to Pollitt et al. (2007), item difficulty consists of the proportion of individuals answering a given question correctly while item demand refers to the cognitive processes necessary for successful completion of a task. In this study, we adopt Pollitt et al.'s (2007) operationalisation of both terms. Item difficulty is the empirical location of an item on a scale. Item demands designate the set of knowledge, comprehension, skills and processes required for a student to respond fully or partially correctly to an item (Ferrara & Duncan, 2011; Ferrara et al., 2011).

The relationship between demands and difficulty is not straightforward (Pollitt et al. 2007; Ferrara et al. 2011). More demanding questions tend to be of greater difficulty while the reverse relationship does not necessarily pertain (Pollitt et al., 2007). Pollitt and colleagues provided a concrete example illustrating how the same maths item administered to the same age group in England and Scotland was much more difficult in England with a 59% success rate compared to 79% in Scotland (Pollitt et al., 2007, pp. 105–107). The authors argued that the cognitive demands were the same for English and Scottish examinees because the item was the same, but due to the difference in curricula, Scottish students were more prepared to answer the question and had a higher rate of success; that is, lower item difficulty. This raises the (positivist) notion that there is an *a priori* psychological demand for each question. But pupils come to a test with a background that prepares them to a lesser or greater extent to answer the questions. Therefore separating the concepts of demand and difficulty empirically is tricky because demand and difficulty are confounded by preparation. By educating pupils, we affect their cognition and therefore the demand of the item. Pupils might be able to retrieve a solution or a method from memory and if they have become very expert at solving

a particular kind of problem, this will have formed part of a cognitive chunk. Education can reduce the requirements for working memory and problem-solving in test-taking.

2. Sources of easiness (SOEs) and sources of difficulty (SODs)

Studies carried out on British high-stakes examinations in different subjects (mathematics, geography, science, French and English) identified features that are associated with the level of difficulty of questions (e.g. Pollitt et al. 1985; Pollitt & Ahmed 1999; 2000; Ahmed & Pollitt 1999; Fisher-Hoch et al. 1997). The features were referred to as ‘sources of easiness’ (SOEs) if they made questions easier to answer or they were called ‘sources of difficulty’ (SODs) if they made questions harder to understand. An example of a SOE could be providing a hint for the correct response and an example of a SOD is the occurrence of double negatives in a question.

Fisher-Hoch et al. (1997) analysed GCSE mathematics and pointed out that although many SODs are specific to mathematics (e.g. arithmetic errors, recalling strategies, spatial representation), many were common to different curriculum subjects. For example, technical terms increased the difficulty of items in mathematics and science. Technical terms refer to subject-specific terms such as ‘bisector’ and ‘vector’ in maths or ‘photosynthesis’ and ‘hydration’ in science. It would be worthwhile investigating whether technical vocabulary contributed to the level of difficulty of KS2 science questions.

Bramley et al. (1998) analysed student scripts of GCSE science papers and identified a number of features making science questions harder including complex vocabulary and grammar, the interference of topic-related knowledge and the amount of knowledge recall required to respond to the question. These are reflected as will be shown later in Ferrara et al.’s (2007, 2011) item response demands framework.

Item writers need to balance SOEs and SODs in items to make sure questions elicit the intended demands they wish to examine (Pollitt et al., 2007). Nevertheless, not all SOEs and SODs in test questions are intended by the examiner and related to the construct being assessed. Research on British GCSEs suggested that some of the features associated with item difficulty were construct-irrelevant (Fisher-Hoch et al., 1997). That is, the SOEs and SODs were not related to the attribute measured, and hence required students’ minds to process additional information (Ahmed & Pollitt, 1999; Pollitt & Ahmed, 1999, 2000). For

instance, presenting superfluous information in problem solving questions would place additional cognitive demands on examinees because they would need to process the additional information and select what is relevant to solve the problem.

Despite the large amount of variables identified as influencing the level of difficulty of items, predicting item difficulty accurately remains a challenge in the field (Hambleton & Jirka, 2006; Pollitt et al., 2007). This is largely due to the complex interaction amongst these variables as well a complicated relationship between item demands and item difficulty (Ferrara et al., 2011; Hambleton & Jirka, 2006; Pollitt et al., 2007). Indeed, Fisher-Hoch et al. (1997) and Bramley et al. (1998) have shown that manipulating SODs in examination questions in a way to make them presumably easier did not always lead to an increase in the proportion of students getting these modified items correctly. For some questions, altering SODs had the opposite effect.

3. Models of item response demands

Pollitt et al. (2007, 1998) described a number of demands that may be placed on examinees when taking a test. These included the duration of the assessment, the amount of work to be carried out, the amount of reading and/or writing required, the readability level of the assessment relative to participants' reading proficiency, demands placed on the working and the long-term memory and the level of stress. Pollitt et al. also proposed the CRAS framework that enabled the judgment of item demands based on five scales: (1) question complexity, (2) resources available, (3) level of abstractness of the question, (4) task strategy and (5) response strategy. This framework is used in the UK by exam boards including Cambridge Assessment¹ to judge demands of exam papers. However, to this date, there is no published study providing empirical evidence supporting the framework's successful application.

Ferrara et al. (Ferrara & Duncan, 2011; Ferrara et al., 2007, 2011) developed a framework of item response demands where they distinguished between cognitive item demands and linguistic item demands. Cognitive item demands referred to reading load, depth of knowledge as described by Webb (Webb, 2007), content complexity and question type. Linguistic demands related to challenging linguistic features in items and included variables such as: ambiguous words, technical vocabulary, complex verbs, pronouns and prepositional

¹ <http://www.cambridgeassessment.org.uk/insights/using-the-cras-framework/>

phrases. Ferrara et al.'s framework (2007, 2011) was employed in two different studies. The first examined the comparability between intended and targeted constructs in science tests (Ferrara & Duncan, 2011) while the second investigated the extent to which item demands in mathematics tests reflected the vertical progression of curricular standards (Ferrara et al., 2007, 2011).

In this paper, we adopted elements of Ferrara et al.'s (2007, 2011) framework along with other variables outlined in the literature to analyse KS2 science items administered over three consecutive years (2010-2012). Expert judges coded items based on a number of independent variables including topic, sub-topic, concept, item type, nature of stimulus, depth of knowledge and linguistic demands. Estimates of item difficulty were generated separately using item response theory (IRT) models. Later, stepwise regression models using the aforementioned independent variables were run to predict relative item difficulty. The next section describes the methods adopted more thoroughly.

Methods

The analysis strategy involved linear regression models to analyse the extent to which selected independent variables predicted the difficulty of KS2 science items estimated from a graded response IRT model. Regression analyses have been carried out previously in this type of research (e.g. Bramley et al., 1998; Ferrara et al., 2011; Fisher-Hoch et al., 1997; Pollitt et al., 1985). This section provides an overview of KS2 science assessments, describes the data-set, presents the dependent as well as the independent variables and outlines the data analysis procedures.

4. Key Stage 2 science assessments

KS2 National Curriculum tests are externally set assessments administered to eleven year olds in England and Wales in core subjects (English, mathematics and science) at the end of primary school (year 6). Until 2012², KS2 science tests have been administered annually to a representative sample of schools³. Matrix sampling and biennial administration has been adopted since 2014. The science test consists of two papers (Test A and Test B). Each paper

² Since 2014, a new matrix sampling design has been adopted with KS2 science assessments only administered biennially.

³ <https://www.gov.uk/government/news/new-video-published-on-changes-to-2016-tests-and-assessments>

is 45 minutes long and has a maximum score of 40. Item formats include objective questions (multiple choice, matching, true or false, etc.) and short constructed questions (one word to a couple of sentences). Most items are dichotomous items (i.e. two response categories 0 and 1), with a few two-mark items.

The purpose of KS2 tests is primarily to provide indicators of national literacy and numeracy standards at the end of primary schooling (Stobart, 2001). The results of KS2 mathematics and English have been used as an accountability index of school performance and a basis to help set educational targets while the outcomes of KS2 science tests provide a measure of national performance in science at the end of primary school (DFEE, 1999; Stobart, 2001). In this respect, student achievement in KS2 is of high significance for schools, local authorities and governments (Stobart, 2001). Decisions about school funding and closure are highly influenced by school performance on these tests. The stakes are however much lower for students despite empirical evidence of pupils experiencing school and teacher pressure to perform highly on them (Reay & Wiliam, 1999; Stobart, 2001).

He et al. (2014) summarised the development process of KS2 science items. Science test items are developed by experienced item writers. The quality of items is reviewed over two phases. First, panels of subject experts including teachers, markers and local authorities judge items in terms of their alignment to the National Curriculum, potential bias and cultural sensitivities. Later, items are pre-tested before being used in live assessment to evaluate their quality further. Item statistics including item difficulty parameters are produced and the appropriateness of items to the target population in terms of their level of difficulty is examined.

5. The dataset

The data analysed in this study consisted of all science items (N=216) administered in two pre-test cycles between 2010 and 2012 and involved an overall sample of 4,164 students in England. The tests were scaled using a two-parameter graded response model (see Samejima, 1997) by the Standards and Testing Agency (STA) and item-level difficulty (thresholds) and discrimination parameters were released to the researchers following an agreed procedure. In addition, the analysis also included the KS2 science specification followed between 2010 and

2012 as well as the respective marking schemes. The science papers, the marking schemes and the KS2 science specification were retrieved from an online public resource⁴.

6. Dependent variable

This study investigated the potential effect of a number of independent variables described below on item difficulty. Given the focus of the study, item discrimination was not included in the analyses. Most KS2 science items were dichotomous and therefore possessed one threshold estimate, that is, one item difficulty estimate. For partial credit items marked out of 2 marks (i.e. items with two threshold estimates), single item difficulty measures were computed by averaging the two threshold estimates for each item.

7. Independent variables

The independent variables considered in this study included: (1) curricular variables (topic, sub-topic and concept), (2) item type, (3) depth of knowledge, (4) nature of the stimulus and (5) language variables.

Curricular variables

Bramley et al. (1998) and Ferrara and Duncan (2011) highlighted the importance of concept complexity in predicting item difficulty in science. Before 2014, the KS2 national science curriculum, referred to as Programme of Study (POS), was organised around three levels in decreasing order of breadth: topic, sub-topic and ‘knowledge, skills and understanding’ (KSU) concept. The programme covered five topics, three of which can be thought of as the conventional scientific subjects: biology (referred to as ‘Life processes and living things’ in the specification), chemistry (described as ‘Materials and their properties’) and physics (termed as ‘Physical processes’). In addition, the specification included two more generic scientific topics: ‘scientific inquiry’ and ‘breadth of study (BoS)’.

Each of these broad scientific topics comprised a number of sub-topics. For instance, ‘Life processes and living things’ includes sub-topics such as ‘life processes’, ‘humans and other animals’, ‘green plants’, ‘variation and classification’, ‘living things in their environment’. Each sub-topic included a number of specific concepts referred to as ‘Knowledge, Skills and Understanding’ (KSU) concepts in the KS2 POS. One of the KSU concepts listed under the sub-topic ‘Life processes and living things’ was ‘that the life processes common to plants

⁴ <http://www.sats-papers.co.uk/sats-papers-ks2.php?route=mustregister>

include growth, nutrition and reproduction'. A total of 81 KSU concepts were described in the KS2 POS⁵.

In this study, we ran regression analyses at topic, sub-topic and KSU concept level to investigate whether curricular variables at any level of specificity predicted item difficulty better. Items were assigned a topic code, one or two sub-topic codes and one or two KSU concepts codes based on the ones specified in the test marking scheme.

Item type

Item type is highly associated with item difficulty (Ferrara & Duncan, 2011; Ferrara et al., 2011; Hambleton & Jirka, 2006). As noted earlier, item type varied in KS2 science tests. It included objective questions and short constructed-response questions. Objective questions included multiple choice questions, matching questions, yes/no or true/false questions and 'complete the table' questions when examinees needed to tick appropriate cells in a table. Given that all non-objective questions were constructed questions that required at most a three-sentence answer and given the age group of the examinees, a distinction was made between short constructed questions that required one word or a simple phrase as an answer from questions requiring longer answers of one to three sentences. For example, item 3b in Test A of series 2010 was considered a short constructed question because it accepted 'thermometer' or 'temperature sensor' as possible answers. Item 3c of the same paper was coded as an extended constructed question because a possible answer was 'wood is a (thermal) insulator'. Occasionally, short constructed questions required examinees to complete a graph or a diagram. In a few instances, extended constructed questions consisted of drawing a schematic representation (e.g. a schematic representation of an electrical circuit). Table 1 below summarises the framework used to code item type.

⁵https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/328904/2014_KS2_science_sample_sample_materials.pdf

Table 1: Framework adopted for coding item type

Type of item	Description
Multiple choice question	Selection of a single response from typically a list of 3 or 4 options.
True or False/ Yes or No questions	Identifying whether a statement is true or false. OR selecting YES or NO as an appropriate response for a question.
Matching	Pairing items in one list of column to items in another list or column.
Complete a table	Tick the correct cells in a table OR add the correct terms/numbers in the cells of a table
Short constructed question	Written response consisting of a number, a term or a short phrase. Occasionally, consists of completing a graph or a diagram.
Extended constructed question	Written response consisting of a least one sentence and up to three sentences. Occasionally, consists of drawing a graph or a diagram.

Depth of knowledge

Ferrara et al. (2011) adopted Webb's (2007) depth of knowledge (DOK) framework to categorise the nature of knowledge and its complexity in items. Based on this framework, questions requiring examinees to recall facts are less demanding than items requiring examinees to develop an argument and support it with evidence. Webb's DOK framework classified knowledge into four levels. In this study, we adopted Webb's DOK framework to code the depth of knowledge in KS2 science items. The framework adopted is presented in Table 2 below. Based on this framework, items requiring retrieving information from rote memory, were coded as DOK1 (Recall), items requiring additional cognitive processing yet not involving use of empirical evidence were coded as DOK2 (Skill/concept), items involving reasoning and making use of evidence were coded as DOK3 (Strategic thinking) and items necessitating sophisticated problem solving skills such as investigation and processing of multiple conditions are coded as DOK4 (Extended thinking).

Table 2: Framework adopted for coding depth of knowledge (DOK) in KS2 science items

DOK level	Title	Description
DOK1	Recall	Recall of information such as a fact, a definition, a simple procedure (e.g. identify, recall, recognise, etc.) <i>e.g. What is the equipment in the picture below called? [KS2 assessment, year 2010, paper A, item 1a]</i>
DOK2	Skill/concept	Engagement in cognitive processing of information and selection of appropriate approaches to solve problems in more than one step (e.g. classifying, organising, comparing, etc.) <i>e.g. Describe the difference in reaction times between the children who play computer games and the children who do not. [KS2 assessment, year 2011, paper A, item 3c]</i>

DOK level	Title	Description
DOK3	Strategic thinking	Engagement in high level reasoning, using evidence, planning a response and explaining thinking (e.g. drawing conclusions from observations, providing evidence, explaining phenomena, etc.) <i>e.g. What is likely to happen to people who drink stream water with micro-organisms in it? [KS2 assessment, year 2010, paper A, item 1d]</i>
DOK4	Extended thinking	Engagement in complex reasoning and elaborate planning of response. Applying significant conceptual understanding and higher order thinking. Building relationships amongst ideas. Developing arguments and supporting them with evidence (e.g. developing and proving hypotheses, designing and conducting experiments, making connections between a finding and relevant concepts, etc.) <i>No examples available in KS2 assessments between 2010 and 2012.</i>

Nature of stimulus

Learning science relies highly on a good command of its language (Norris & Phillips, 2003; Osborne, 2002). Indeed, recent approaches to science education focus on scientific literacy and emphasise the importance of acquiring the scientific language for high performance in science. The scientific language consists of a combination of texts, mathematical equations and visual representations such as pictures, graphs, tables, maps, etc. (Anagnostopoulou, Hatzinikita, & Christidou, 2012; Norris & Phillips, 2003; Osborne, 2002). New science assessments comprise textual materials as well as visual illustrations. For instance, science items of the Programme for International Student Assessment (PISA) are organised around one or two stimuli which could include a text or visual materials such as picture, graph or table (OECD, 2009).

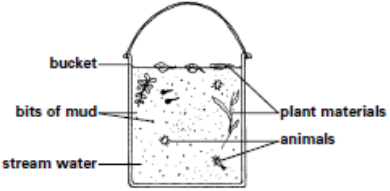
KS2 science assessments are no different. KS2 questions are also organised around one or a couple of stimulus materials (text, photo, table, graph, schematic representation). The figure below shows an example of KS2 science items. Unit 1 ‘Drinking water’ of the 2010 series (Test A) includes six items (*a* to *f*). Figure 1 below shows two items only due to space restrictions (items 1a and 1b). The two pictorial representations are stimuli and coded as drawings based on an adaptation of Moline’s (1995) framework which categorises visual representations into various types (e.g. photo, drawings, flowchart, graphs, tables). Table 3 presents the adapted version of Moline’s framework. Moline’s framework has been applied

in Anagnostopoulou et al.'s study (2012) to compare the nature of visual representations occurring in PISA science assessments with the ones included in Greek biology textbooks.

Figure 1: Example of a KS2 science item

1 Drinking water

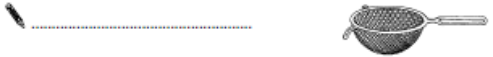
(a) People who walk in the mountains can travel a long way from towns. They might have to get their drinking water from a stream. The water must be made safe before they drink it.



bucket
bits of mud
stream water
plant materials
animals

Animals can be separated from the water using the equipment shown below.

What is the equipment in the picture below called?



(1 mark)

(b) It is important to put the animals back where they were found.

Write **true** or **false** next to each statement to show why it is important for the animals to be put back in the stream.

True or false?

because the animals are adapted to live in the stream
so the animals do not get eaten by predators
(1 mark)

Source: KS2 Test A (2010 series)

Table 3: Framework adopted for coding stimuli in KS2 science tests

Stimulus	Description
Photo	Picture of a real object/animal/person produced by a camera
Drawing	A picture that is sketched without the use of a camera
Schematic representation	A representation of a mechanism or process using scientific conventions (e.g. electrical circuit)
Graph	Information plotted on a graph, histogram, pie chart, etc.
Table	Tabulated information
Flowchart	A chart showing a logical progression

Stimulus	Description
Text	Textual material (e.g. newspaper excerpt, advertisement)

Various theories in psychology such as dual coding and cognitive load theory supported the combination of verbal and visual inputs to foster meaningful learning (Levie & Lentz, 1982; Mayer & Moreno, 2003; Mayer, 1997) and as a tool to assess deep understanding and knowledge transfer (Yore & Treagust, 2006). The effect of illustrations on the understanding of instructional texts has been extensively explored (e.g. Carney & Levin, 2002; Filippatou & Pumfrey, 1996; Levie & Lentz, 1982). Research suggests that the effect of pictures on text comprehension depends largely on student characteristics as well as the nature of the illustrations, their function (e.g. decorative, representational, organization, etc.) and the level of difficulty of the text. While decorative pictures did not lead to any benefits in understanding textual material and sometimes resulted in confusion, illustrations that genuinely represented the text and organised ideas within it increased its comprehension (Carney & Levin, 2002). In this study, decorative illustrations were excluded from the analysis as they could not be associated with particular items.

Shorrocks-Taylor and Hargreaves (1999) queried the role of illustrations in KS2 mathematics assessments and argued that the relationship between mathematics items and pictures might be comparable to the one described above (i.e. textual material and illustrations). The illustrations examined in the literature referred mainly to pictorial representations but did not explicitly include graphs and tables. Anagnostopoulou et al. (2012) argued that the limited familiarity of Greek students with the types and function of visual illustrations used in PISA science assessment may be a source of their low achievement on this test.

Language variables

Evidence of the impact of language on item difficulty in science is ample (e.g. Bramley et al., 1998; Fisher-Hoch et al., 1997; Pollitt et al., 1985) despite some research reporting insignificant effect (e.g. Ferrara & Duncan, 2011). In this study, language demands have been investigated using *Coh-Metrix*, an open computational linguistic facility grounded in theory of learning and discourse comprehension (Graesser, McNamara, & Kulikowich, 2011). The use of *Coh-Metrix* has been highly recommended for teachers to evaluate complexity and appropriateness of texts used in schools (Crossley et al., 2007; Crossley et al., 2008). However, no prior published research has employed this software for the analysis of linguistic demands in assessments. Using an automated computational facility has several

advantages such as time efficiency and limiting bias due to human judgement that may create inconsistencies in the coding of variables.

Coh-Metrix generates over a hundred linguistic indicators categorised into five dimensions: (1) narrativity; (2) syntactic simplicity; (3) word concreteness; (4) referential cohesion; (5) deep cohesion. The five dimensions, described in Table 4 below, have been shown by Graesser et al. (2011) to contribute to text difficulty. The first three dimensions have been identified in the literature reviewed earlier (e.g. Bramley et al., 1998; Pollitt et al., 2007, 1985) yet referential and deep cohesion have not been investigated in educational assessments. In a study in Germany comparing the performance of second language learners (SLL) in mathematics to that of native speakers, items favouring SLL were found to be more cohesive (Haag et al., 2013). In addition to these five dimensions, we reported sentence length (i.e. word count per sentence) and paragraph length (sentence count per paragraph) generated by *Coh-Metrix* for some descriptive statistics.

Table 4: The five dimensions of Coh-Metrix

Dimension	Description
Narrativity	Extent to which the item uses language comparable to everyday language
Syntactic simplicity	The degree to which the item is concise and makes use of simple and familiar syntactic structures
Word concreteness	The degree to which the vocabulary use is concrete and meaningful
Referential cohesion	The degree of overlap of words and ideas across sentences forming explicit connections
Deep cohesion	The extent to which the item contains causal and intentional connectives (e.g. because) that help the reader build connections and understand relationships and processes in the text

8. Coding of items

Items were coded based on the independent variables described earlier and the corresponding coding frameworks. For instance, item 1a of Figure 1 above was coded as a short constructed question as it required examinees to provide the name of the equipment in the picture (*sieve*). The visual representation of the sieve was a sketch that was not generated using a camera nor

did the picture depict a mechanism; it was a simple *drawing* of the filtering tool. The POS of KS2 included the example of sieve as equipment permitting the separation of big elements from smaller ones. This meant that examinees had been introduced to this particular tool in their science class. Question 1a was rated at DOK level 1 because examinees had to only *identify* the equipment in the picture and *recall* its name.

Curricular variables were coded using marking schemes which attributed one or two specific KSU concepts to each item. Given that KSU concepts were related to a specific topic and sub-topic, all levels of the curricular variables were coded for each item (that is, topic, sub-topic and KSU concept). Question 1a was accordingly coded as a scientific enquiry (topic), planning (sub-topic) item, which required students to ‘think about what might happen or try things out when deciding what to do, what kind of evidence to collect, and what equipment and materials to use⁶’. Estimates of the language variables have all been generated using *Coh-Metrix* software (Graesser et al., 2011). For question 1a, the following text was included in the analysis:

People who walk in the mountains can travel a long way from towns.
They might have to get their drinking water from a stream.
The water must be made safe before they drink it.
Animals can be separated from the water using the equipment shown below:
What is the equipment in the picture below called?

Coh-Metrix generated the values for the following variables: sentence count per item, word count per item, narrativity, syntax simplicity, concreteness, referential cohesion and deep cohesion.

With the exception of language variables, the coding of all the other variables was carried out independently by the first author and a PhD student who is also an experienced item developer. Both raters had expert knowledge in science and were based at a university in England. The first author described and discussed the rating framework with the second rater before the coding of items. Consensus meetings were carried out following the coding phase to debate disagreements in the rating of some items. The percentage of rater agreement was calculated for these variables.

9. Data analysis

Data analysis consisted of running stepwise regression analyses to assess the amount of variance explained in item difficulty (dependent variable) based on the independent variables.

⁶ The POS for KS2 can be found on <http://www.sats-papers.co.uk/sats-papers-ks2.php?route=mustregister>

Descriptive statistics were computed for various variables and linear regression analyses were run. Regression analyses have been commonly used in this type of research (e.g. Bramley et al., 1998; Ferrara et al., 2011; Fisher-Hoch et al., 1997; Pollitt et al., 1985). We carried out stepwise regression analyses because they are helpful in progressively selecting predictors from a large number of explanatory variables when there is not a strong theory for which variables should contribute to the prediction. Two approaches were used: forward selection and backward elimination. In the forward selection approach, the starting point included no variables and potential explanatory variables were added progressively while in the backward elimination, the analysis started with all possible explanatory variables and variables were deleted gradually until no further improvement was possible. At each step, an F -statistic was generated for the estimated coefficient of each variable included (or excluded in the backward elimination models). Model improvement was examined at each step by comparing the generated F -statistic with a model comparison criterion, an F -statistic threshold (see Tabachnik & Fidell, 2007). In both approaches, the model comparison criterion was $F \leq 0.050$.

Results

This section describes the results of the regression analyses following some descriptive statistics of the depended and independent variables.

10. Descriptive statistics

This section presents some descriptive statistics of the dependent and independent variables necessary for the interpretation of results.

Dependent variable: Item difficulty

Table 5 below provides some descriptive statistics of item difficulty estimates of the 216 KS2 science items analysed. The results suggest that the items were easy for the population of test-takers, as the mean item difficulty is -0.73 with little variability around it ($SE=0.09$) and the distribution shows a clear skew to the left indicating that the science items of the series 2010, 2011 and 2012 were too easy for the sample tested. This is consistent with He et al.'s findings (2014) on the 2010 and 2011 series.

Table 5: Descriptive statistic of item difficulty estimates (N=216)

	Statistic (SE)
Mean	-0.73 (0.09)
Skewness	-1.225 (0.17)
Kurtosis	6.37 (0.33)
Maximum (hardest item)	-8.51
Minimum score (easiest item)	3.09

Independent variables

Tables 6 to 10 provide descriptive statistics of the independent variables: curricular variables, item type, DOK, nature of stimulus and language variables.

Curricular variables

The frequencies of topics, sub-topics and KSU concepts were examined. ‘Scientific enquiry’ was the topic with the highest frequency (45% of items) whereas ‘breadth of study’ had the lowest frequency (1.4% items). There was a balance in the coverage of the three topics reflecting the traditional scientific disciplines (biology, chemistry and physics) with a frequency ranging between 20% and 21%. Given the low variability in the topic variable amongst the three core topics, we will not be discussing this variable any further.

At sub-topic level, the frequency of variables ranged between 3% and 10% of items except for two concepts: ‘knowledge, skills, concept’ did not appear in any item while ‘investigating skills’ appeared in almost half of the items (42.1%).

Table 6: Descriptive statistics of ‘Sub-topic’ variable (N=216)

Topic	Sub-topic	Frequency	Percentage (%)
Scientific enquiry	Ideas and evidence in science	6	2.8
	Investigating skills	91	42.1
Life processes and living things	Life processes	6	2.8
	Humans and other animals	10	4.6
	Green plants	13	6.0
	Variation and classification	7	3.2
	Living things in their environment	13	6.0
Materials and their properties	Grouping and classifying materials	14	6.5
	Changing materials	20	9.3
	Separating mixtures of materials	18	8.3
Physical properties	Electricity-simple circuits	7	3.2
	Types of forces	13	6.0
	Light and sound	16	7.4
	The Earth and beyond	7	3.2

Breadth of study	Knowledge, skills & understanding	0	0
	Communication	3	1.4

Descriptive statistics for the KSU concept variables were examined. However, given the large number of these variables (n=81) and the limited space, we will only provide a brief overview of the results. In general, the frequency of each KSU concept was between 2 to 5, with some KSU concepts being completely absent and a few KSU concepts (under the investigative skills sub-topic) featuring in as many as 19 items (e.g. KSU 5: use simple equipment and materials appropriately and take action to control risks).

Coders verified that their codes matched with the topic, sub-topics and KSU concepts specified in the marking scheme for each item. Given that the codes were not based on expert judgement, inter-coder agreement was not computed for this variable.

Item type

The rate of inter-coder agreement for item type was 69% with up to 87% agreement in judging the type of objective questions. There was a lower agreement in judging short constructed and extended constructed questions due to the way some answers were expressed in the marking scheme. Some responses were formulated using a sentence while the answer could be easily reduced to a short phrase. For instance in Test B of the 2010 series, item 4c accepted two responses: ‘vibrations’ and ‘she can feel the floor vibrate’. Such an item was coded as short constructed response question. The relatively high inter-coder agreement (cf. Ferrara et al., 2011) suggests that the framework of coding adopted worked fairly well. Table 7 below shows the frequency of item type in KS2 science tests of the 2010, 2011 and 2012 series. Based on the coding framework for item type described earlier, almost half of the questions appeared to be short constructed questions (41%) and a just over a fourth consisting of extended constructed questions (26%). Objective questions accounted for only a third of the total number of items (33%) with multiple choice items being the most frequently used (23%).

Table 7: Descriptive statistics of ‘Item type’ variable

Item type	Frequency	Percentage (%)
Multiple choice	49	22.7
True/False and Yes/No	15	6.9
Matching	5	2.3
Complete table	2	0.9
Short constructed	88	40.7
Extended constructed	57	26.4
N	216	100.0

Depth of knowledge

The rate of inter-coder agreement for depth of knowledge was low (30%). Indeed, during consensus meetings raters expressed challenges in applying Webb's (2007) framework to judge the depth of knowledge required to answer KS2 questions. Table 8 below summarises the frequency of DOK levels in KS2 items. More than half of the items (62%) were at the lowest DOK level 'recall and reproduce'. More than a third targeted the second level of DOK, 'skills and concept' (35%). Only 8 items (less than 4%) involved strategic thinking (DOK 3) and no question targeted the highest DOK level (extended thinking). These findings are not surprising given the age group and the overall level of difficulty of the tests.

Table 8: Descriptive statistics of 'Depth of Knowledge' variable

Depth of Knowledge	Frequency	Percentage (%)
Recall & reproduce (DOK1)	133	61.6
Skills & concept (DOK2)	75	34.7
Strategic thinking (DOK3)	8	3.7
Extended thinking (DOK4)	0	0
N	216	100

Nature of stimulus

The rate of inter-coder agreement for this variable was satisfactory (55%). Discrepancy in the judgement arose between drawing and schematic representation because of the vague way 'schematic representation' was described in the initial framework (i.e. a representation of a mechanism or process). The raters then agreed to add the phrase 'using scientific conventions' to make the description less vague.

Table 9 below suggests that most commonly used stimulus consisted of drawings (21%). Photo and tables were also frequently used (19% and 14% respectively) with much less use of schematic representations and graphs (7.4% and 5.1% respectively). This finding was not surprising given the age group of the students and the overall level of difficulty of KS2 items as outlined earlier. Photos and drawings are normally much easier to process than graphs and schematic representations (cf. Sharma, 2006).

Table 9: Descriptive statistics of 'Stimulus' variable

Stimulus type	Frequency	Percentage (%)
Photo	41	19
Drawing	45	20.8
Schematic representation	16	7.4
Graph	11	5.1
Table	31	14.4
Flowchart	2	0.9
Text	5	2.3

N

216

100

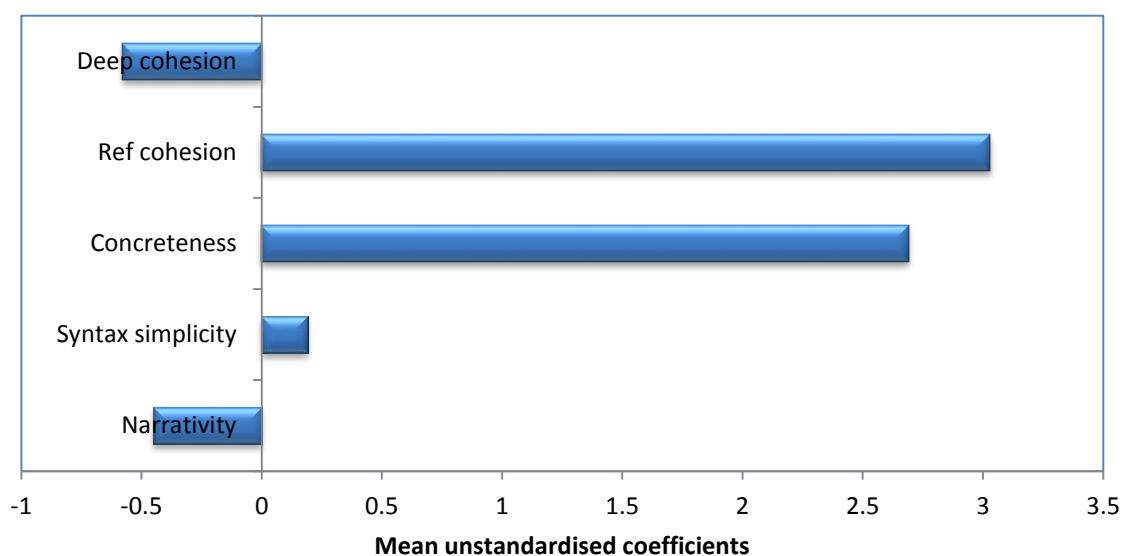
Language variables

Language variables analysed using *Coh-Metrix* (Graesser et al., 2011) yielded the results summarised in Table 10 below. On average, items consisted of 4 sentences with around 40 words. Negative narrativity of -0.45 suggests that the items are written in formal language that does not resemble everyday language. This is supported by the low syntax simplicity indicator of 0.19 which points to a high syntax complexity. The mean concreteness index was 2.69 indicating a very low level of abstraction. Referential cohesion was 3.03. This value is considered very high indicating a great overlap between ideas across sentences within an item. Nevertheless, deep cohesion was negative (-0.58) pointing to very few connectives between sentences making it more demanding for students to deduce relationships between elements within a question.

Table 10: Descriptive statistics of 'Coh-Metrix' language dimensions (N=216)

	SC/item	WC/ item	Narrativity	Syntax simplicity	Concreteness	Ref cohesion	Deep cohesion
Mean	4.14	39.38	-.45	.19	2.69	3.03	-.58
SE	.17	1.34	.067	.089	.12	.15	.14

Figure 2: Five dimensions of Coh-Metrix



The results detailed in this section provide a context for the stepwise regression models described in the next section. It is worth noting that some of the coding frameworks adopted

worked better than others. This has great implications on the power of the predictive models, a point which would be discussed more elaborately in the discussion section.

11. Regression models

Stepwise linear regression models were run at sub-topic and KSU concept levels to investigate the amount of variance predicted at each level of curricular variable. As noted earlier, given the breadth and the distribution of topics, regressions at topic level were not further examined.

Sub-topic level

The stepwise linear regression analysis produced similar models; with the forward model being virtually a subset of the backward model. We will only be discussing the results of the forward model to avoid over-estimation of model prediction. At sub-topic level, 16% of the variance in item difficulty was explained by five independent variables: extended constructed questions, photo, sub-topics 4, 5 and 7 ($R^2 = 0.16$).

Table 11 below suggests that an extended constructed question increased the difficulty of the item by 0.88 units keeping all other independent variables constant. Including a photo in an item decreased its difficulty by 0.44 units, all other variables kept fixed. When an item targeted sub-topic 5 (Green plants), the difficulty of the item increased by 0.76 units, all other variables kept constant. When an item targeted either sub-topic 4 (Humans and other animals) or sub-topic 7 (Living things in their environment), the difficulty of the item decreased respectively by 1.18 and 0.66 units (all other variables kept constant). It is worth noting that sub-topics 4, 5 and 7 are all related to a single topic: Living processes and living things in the KS2 POS.

Table 11: Coefficients of the forward model (sub-topic level)

	Freq (items)	Coefficients	SE	T	Sig
(Constant)		-0.829	0.102	-8.094	0.000
Extended constructed	57	0.881	0.178	4.936	0.000
Humans and other animals (Sub-T4)	10	-1.177	0.371	-3.172	0.002
Photo	41	-0.443	0.199	-2.221	0.027
Green plants (Sub-T5)	13	0.755	0.329	2.298	0.023
Living things in their environment (Sub-T7)	13	-0.660	0.330	-2.002	0.047

KSU concept level

The multiple regression analysis produced two models with one, the forward regression model ($R^2 = 0.22$), being a subset of the other, the backward regression model ($R^2 = 0.25$).

Here again, we will only discuss the forward model to avoid over-estimating the amount of variance predicted.

The forward regression model identified nine significant variables. Table 12 below suggests that apart from one variable – Extended constructed response – the other eight variables were KSU concept variables (see Table 13 for a description). Consistent with the sub-topic level analysis, extended constructed question increased the level of difficulty of items (all other variables kept constant) while photo decreased the level of difficulty of items. Some KSU concept variables decreased the level of difficulty of items (KSU7, KSU24, KSU25, KSU36) while others increased it (KSU30, KSU45, KSU73).

Table 12: Coefficients of the forward model (at KSU concept level)

	Freq (items)	Coefficients	SE	T	Sig
(Constant)		-0.884	0.097	-9.146	0.000
Extended constructed	57	0.983	0.173	5.685	0.000
KSU7	3	-1.740	0.640	-2.720	0.007
KSU24	2	-2.006	0.782	-2.566	0.011
Photo	41	-0.468	0.193	-2.422	0.016
KSU25	2	-1.852	0.782	-2.369	0.019
KSU73	2	1.497	0.641	2.336	0.020
KSU30	4	1.238	0.556	2.229	0.027
KSU45	3	1.423	0.641	2.222	0.027
KSU36	3	-1.395	0.642	-2.172	0.031

Table 13: Description of KSU concept predictors

KSU	Description
KSU7	use simple equipment and materials appropriately and take action to control risks
KSU24	about the main stages of the human life cycle
KSU25	about the effects on the human body of tobacco, alcohol and other drugs, and how these relate to their personal health
KSU30	about the parts of the flower [for example, stigma, stamen, petal, sepal] and their role in the life cycle of flowering plants, including pollination, seed formation, seed dispersal and germination
KSU36	how animals and plants in two different habitats are suited to their environment
KSU45	to describe changes that occur when materials are mixed [for example, adding salt to water]
KSU73	how the position of the Sun appears to change during the day, and how shadows change as this happens

Concepts KSU30, KSU45 and KSU73 have been identified in the science education literature as being challenging for primary students (Barker, 2000; Jones, Lynch, & Reesink, 1987; Kind, 2004; Papageorgiou & Sakka, 2000; Parker & Heywood, 1998;). KSU30 relates to the growth and development of plants. Lin (2004) noted that this concept has not been widely researched although it is one that merits some attention. According to the author, the concept of growth and development is particularly demanding for children as it includes a morphological aspect and a biochemical physiological aspect. Hence, students need not only understand the organism's anatomy and its changes as a result of growth. They also need to comprehend all the physiological processes and chemistry involved. Items relating to such a complex concept are likely to be more demanding and may hence be of higher difficulty. KSU45 relates to changes observed in mixtures and solutions. Kind (2004) reported in a review prepared to the Royal Society of Chemistry that students starting post-16 chemistry courses may exhibit poor understanding of fundamental concepts in chemistry such as the distinction between elements, mixtures and compounds. It is not surprising then to see that students find these concepts demanding at the end of primary school especially when in many cases, these elementary concepts appear to be misconceived by their teachers (e.g. Papageorgiou & Sakka, 2000). These demands are likely to be manifested in KS2 science items and increase the proportion of students responding incorrectly to these items. KSU73 deals with the relationship between the position of the sun *vis-à-vis* planet Earth and the appearance of day and night. Concepts related to astronomical events have been reported as exceptionally challenging for students (Barker, 2000; Jones et al., 1987; Sharp, 1996; Vosniadou, 1991). Children often develop alternative astronomical models that do not conform to scientifically accepted models (Jones et al., 1987). Parker and Heywood (1998) reported that pre-service teachers needed particular support in terms of teacher subject knowledge for that particular KS2 unit (Topic 4: *The Earth and beyond*). Teachers' poor mastery of a concept may add to the inherent demand of the concept and potentially further increase the difficulty of relevant items in science tests. While the amount of variance predicted by the model at KSU concept level was consistent with previous research (e.g. Ferrara & Steedle, 2014), the low frequency of significant KSU concepts in the dataset limits the amount of generalisation and inferences one can make based on the study. KSU30 appears in four items only, KSU45 occurs in three items only while KSU73 figures in two items only. Considering the results of regression models at sub-topic level appeared to be more reasonable to generalise. Unfortunately however, only 16% of the

variance was explained using this model. This means there is still 84% variance unexplained indicating great room for improvement.

Discussion

Educational assessment rests on the assumption that students' understanding and skills can be measured by exposing them to questions that elicit specific cognitive processes in their minds. Hence, it is crucial that test writers control what makes items more or less challenging and eliminate construct-irrelevant difficulty as it constitutes a major source of question invalidity (Messick, 1993). In this study we identified a number of variables from the literature associated with the level of difficulty of science items. These were used as independent variables in regression models which predicted between 17% and 22% of variance in item difficulty of KS2 science tests. Item type (extended constructed questions) was the highest predictor of item difficulty, significantly increasing the level of difficulty of KS2 science items. Conversely, photos had an opposite effect. Photos were associated with a significantly decreased level of difficulty of KS2 science items. This finding is consistent with research on multimedia instruction which suggests that visual material relieves cognitive load placed on participants' minds when processing textual material (Mayer & Moreno, 2003; Mayer, 2001).

The regression models also pointed to the significance of some curricular variables in explaining the variance. According to the literature in science education, concepts related to 'reproduction in flowering plants', 'mixtures and solution' and 'astronomy' are challenging for students at all school levels. They are even confusing for primary school teachers. For such concepts, item difficulty is not only affected by the potential inherent complexity of the concept but also by the quality of teaching. A limitation in this research was the number of items reflecting some of the KS2 concepts. A larger dataset comprising more items may solve this issue and such a study might help to move the field on generally.

DOK levels did not contribute to the level of difficulty of items. One explanation could be that almost all items targeted low DOK levels (1 and 2). A dataset with more DOK variability might lead to clearer results. Another explanation might be attributed to the framework used. Raters experienced challenges adapting Webb's (2007) framework to KS2 items because the variable was holistic and involved a great level of subjectivity let alone the fact that the framework has been developed in an American assessment context. Challenges in judging

item demands of questions using pre-set frameworks has been reported to be problematic in the literature (Hambleton & Jirka, 2006; Pollitt et al., 2007). A great amount of training should precede the judgement activity and perhaps a method inspired by Hambleton and Jirka's (2006) study where judges are provided with items with known item demand levels to use as anchors to make judgement about other items.

Language variables did not seem to contribute to the level of item difficulty. This is likely to be due to the extent to which computational linguistic facilities are less effective with very short textual materials. This is due to the fact that these techniques typically require more textual context to infer information about the nature of an item, whereas humans are able to infer from their broad world knowledge. As technology advances, computational linguistic software may be able to handle short texts as well as deeper levels of semantic information (e.g., Landauer, Foltz & Laham, 1998) to provide more conclusive results.

It is worth noting that the two most significant predictors of item difficulty in this study were related to question format (item type) and presentation (visual representation) and were not directly related to the scientific constructs measured (i.e. scientific concept and depth of knowledge). These results and the low percentage of variance explained in this study do not necessarily indicate that the test is of poor validity. Such conclusions need additional and more thorough data.

Predicting item difficulty has been tormenting test developers for a long time. The unpredictable interaction of variables is certainly one of the reasons this area is extremely challenging. This is probably the case with KS2 science tests too. The literature points to particularly challenging scientific concepts in school curricula. By modifying some of the various variables (e.g. language, type of item, nature of the stimulus, etc.), one can theoretically write easier or harder questions on the same topic. However, we are still incapable of predicting item difficulty accurately or determining its direction after its manipulation. This lack of knowledge stems from poor theoretical models of item difficulty and has great implications on pedagogy because often assumptions about students' understanding are based on their performance in tests. In this paper, we have highlighted main methodological limitations for building predictive models. We believe that a way forward is to provide evidence of the strengths and limitations of coding frameworks and reveal weaknesses of rating processes in order to improve them.

Acknowledgements

This project has been funded by Pearson, Inc. The authors are grateful for Dr Barbara Donahue and Mrs Louise Benson at the Standards and Testing Agency for providing item information and parameters as well as their invaluable feedback on a previous draft of the manuscript. The authors would also like to thank Miss Nia Dowell (PhD student at the Institute of Intelligent Systems, University of Memphis) for her assistance with *Coh-Metrix* software and Mrs Diana Ng (DPhil candidate at the Oxford University Centre for Educational Assessment) for participating in the coding of items.

References

- Ahmed, A., & Pollitt, A. (1999). Curriculum demands and question difficulty. *Paper Presented at the International Association for Educational Assessment*. Bled, Slovenia.
- Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: an experimental investigation of focus. *Assessment in Education: Principles, Policy & Practice*, 14(2), 201–232. doi:10.1080/09695940701478909
- Anagnostopoulou, K., Hatzinikita, V., & Christidou, V. (2012). PISA and biology school textbooks: The role of visual material. *Procedia - Social and Behavioral Sciences*, 46, 1839–1845. doi:10.1016/j.sbspro.2012.05.389
- Barker, V. (2000). *Beyond Appearances: Students' misconceptions about basic chemical ideas. A report prepared for the Royal Society of Chemistry*. London: London: Education Division, Royal Society of Chemistry. Retrieved from <http://www.dougdela matter.com/website1/science/philosophy/articles/royal.pdf>
- Black, P., & Wiliam, D. (1999). *Assessment for Learning. Beyond the Black Box*. Assessment Reform Group. Retrieved from <http://www.assessment-reform-group.org/publications.html>.
- Bramley, T., Hughes, S., Fisher-Hoch, H., & Pollitt, A. (1998). *Sources of Difficulty in Examination Questions: Science*. Research and Evaluation Division U.C.L.E.S.
- Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review*, 14(1), 5–26.
- Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3), 475 – 493. doi:10.1002/j.1545-7249.2008.tb001
- Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal*, 91(2), 15–30. doi:10.1111/j.1540-4781.2007.00507.x
- DFEE. (1999). *Weighing the Baby: The Report of the Independent Scrunity Panel on the 1999 Key Stage 2 National Curriculum tests in English and Mathematics*. London: DFEE.
- Ferrara, S., & Duncan, T. (2011). Comparing science achievement constructs: Targeted and

- achieved. *The Educational Forum*, 75(2), 143–156.
- Ferrara, S., Phillips, G., Williams, P., Leinwand, S., Mohoney, S., & Ahadi, S. (2007). Vertically articulated performance standards: An exploratory study of inferences about achievement and growth. In R. Lissitz (Ed.), *Assessing and Modeling Cognitive Development in School* (pp. 31–63). Maple Grove, MN: JAM Press.
- Ferrara, S., & Steedle, J. (2014). *GED Item difficulty modeling results and recommendations. Paper presented to the General Educational Development Testing Service (GEDTS)*. Pearson Center for Next Generation Learning and Assessment.
- Ferrara, S., Svetina, D., Skucha, S., & Davidson, A. H. (2011). Test development with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practice*, 30(4), 3–15. doi:10.1111/j.1745-3992.2011.00218.x
- Filippatou, D., & Pumfrey, P. D. (1996). Pictures, titles, reading accuracy and reading comprehension: a research review (1937-1995). *Educational Research*, 38(3), 259–291.
- Fisher-Hoch, H., Hughes, S., & Bramley, T. (1997). What makes GCSE examination questions difficult? Outcomes of manipulating difficulty of GCSE questions. *Paper Presented at the British Educational Research Association Annual Conference*. University of York: xxx.
- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223–234. doi:10.3102/0013189X11413260
- Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction*, 28, 24–34. doi:10.1016/j.learninstruc.2013.04.001
- Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. . Downing & T. M. . Haladyna (Eds.), *Handbook of test Development* (pp. 399–420). Mahwah, NJ: Lawrence Erlbaum.
- Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. . Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 399–420). Mahwah, NJ: Lawrence Erlbaum Associates.
- He, Q., Anwyll, S., Glanville, M., & Opposs, D. (2014). An investigation of measurement invariance of the Key Stage 2 National Curriculum science sampling test in England. *Research Papers in Education*, 29(2), 211–239. doi:10.1080/00131881.2013.844942
- Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35(1), 69–81.
- Jones, B. E. (1993). *GCSE inter-group cross-moderation studies 1992. Summary report on studies undertaken on the summer 1992 examinations in English, mathematics and science*. Inter-Group Research Committee for the GCSE.
- Jones, B. L., Lynch, P. P., & Reesink, C. (1987). Children's conceptions of the earth, sun and moon. *International Journal of Science Education*, 9(1), 43–53. doi:10.1080/0950069870090106

- Kind, V. (2004). *Beyond Appearances: Students' Misconceptions About Basic Chemical Ideas* (2nd ed.). Durham: Durham University. Retrieved from http://community.nsee.us/pd/pd2007_assessment/misconceptions/Beyond-appearances.pdf
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284
- Levie, W. H., & Lentz, R. (1982). Effects of text illustrations : Review of research. *Educational Communication and Technology*, 30(4), 195-232.
- Lin, S. (2004). Development and application of a two-tier diagnostic test for high school students' understanding of flowering plant growth and development. *International Journal of Science and Mathematics Education*, 2, 175-199.
- Maughan, S., Styles, B., Lin, Y., & Kirkup, C. (2012). Partial estimates of reliability : Parallel form reliability in the Key Stage 2 science tests. In D. Opposs & Q. He (Eds.), *Ofqual Reliability Compendium* (pp. 67-90). Coventry: Ofqual.
- Mayer, R. E. (1997). Multimedia learning : Are we asking the right questions ? *Educational Psychologist*, 32(1), 1-19. doi:10.1207/s15326985ep3201
- Mayer, R. E. (2001). *Multimedia Learning*. New York: Cambridge University Press.
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38, 43-52. doi:10.1207/S15326985EP3801
- McLone, R. R., & Patrick, H. (1990). *Standards in Advanced level mathematics. Report of study 1: A study of the demands made by the two approaches to "double mathematics"*. An investigation conducted by the Standing Research Advisory Committee of the GCE Examining Boards. Cambridge: University of Cambridge Local Examinations Syndicate.
- Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan Publishing Co.
- Meyer, L. (2003). Repeat of AQA Standards Unit analyses originally run for the GCE Economics awarding meeting.
- Moline, S. (1995). *I See What You Mean*. York, ME: Stenhouse.
- Newton, P. E. (2009). The reliability of results from national curriculum testing in England. *Educational Research*, 51(2), 181-212.
- Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, 87(2), 224-240.
- OECD. (2009). *PISA Data Analysis Manual* (Vol. 2nd). France: Paris: OECD.
- Osborne, J. (2002). Science without literacy: A ship without a sail? *Cambridge Journal of Education*, 32(2), 203-218. doi:10.1080/03057640220147559
- Papageorgiou, G., & Sakka, D. (2000). Primary school teachers' views on fundamental chemical concepts. *Chemistry Education: Research and Practice in Europe*, 1(2), 237-

- Parker, J., & Heywood, D. (1998). The earth and beyond: Developing primary teachers' understanding of basic astronomical events. *International Journal of Science Education*, 20(5), 503–520. doi:10.1080/0950069980200501
- Pollitt, A., & Ahmed, A. (1999). A new model of the question answering process. *Paper Presented at the International Association for Educational Assessment*. Bled, Slovenia.
- Pollitt, A., & Ahmed, A. (2000). Comprehension failures in educational assessment. *Paper Presented at the European Conference on Educational Research*. Edinburgh.
- 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, & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 166–206). London: Qualifications and Curriculum Authority.
- Pollitt, A., Entwistle, N. J., Hutchinson, C. J., & De Luca, C. (1985). *What makes exam questions difficult?* Edinburgh: Scottish Academic Press.
- Pollitt, A., Hughes, S., Ahmed, A., Fisher-Hoch, H., & Bramley, T. (1998). *The effects of structure on the demands in GCSE and A level questions*. Report to Qualifications and Curriculum Authority. University of Cambridge Local Examinations Syndicate.
- Reay, D., & Wiliam, D. (1999). 'I' ll be a nothing': structure, agency and the construction of identity through assessment. *British Educational Research Journal*, 25(3), 343–354. doi:10.1080/0141192990250305
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85–100). New York: Springer Science.
- Sharma, S. V. (2006). High school students interpreting tables and graphs: Implications for research. *International Journal of Science and Mathematics Education*, 4(2), 241–268. doi:10.1007/s10763-005-9005-8
- Sharp, J. G. (1996). Children's astronomical beliefs: a preliminary study of Year 6 children in south-west England. *International Journal of Science Education*, 18(6), 685–712.
- Shorrocks-Taylor, D., & Hargreaves, M. (1999). Making it clear: a review of language issues in testing with special reference to the National Curriculum mathematics tests at key stage 2. *Educational Research*, 41(2), 123–136. doi:10.1080/0013188990410201
- Stobart, G. (2001). The validity of National Curriculum assessment. *British Journal of Educational Studies*, 49(1), 26–39. doi:10.1111/1467-8527.t01-1-00161
- Tabachnik, B., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Boston, MA: Pearson.
- Trumper, R. (2001). A cross-age study of junior high school students' conceptions of basic astronomy concepts. *International Journal of Science Education*, 23(11), 1111–1123. doi:10.1080/09500690010025085
- Trumper, R. (2003). The need for change in elementary school teacher training - a cross-

- college age study of future teachers' conceptions of basic astronomy concepts. *Teaching and Teacher Education*, 19, 309–323. doi:10.1080/0013188970390204
- Vosniadou, S. (1991). Designing curricula for conceptual restructuring: lessons from the study of knowledge acquisition in astronomy. *Curriculum Studies*, 23(3), 219–237.
- Webb, N. L. (2007). Issues related to judging the alignment of the curriculum standards and assessments. *Applied Measurement in Education*, 20(1), 7–25.
- William, D. (2001). Reliability, validity, and all that jazz. *Education*, 29(3), 17–21.
- Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education*, 8(4), 341–351.
- Yore, L. D., & Treagust, D. F. (2006). Current Realities and Future Possibilities: Language and science literacy—empowering research and informing instruction. *International Journal of Science Education*, 28(2-3), 291–314. doi:10.1080/09500690500336973