PREDICTIVE MODELS OF QUESTION DIFFICULTY

A critical review of the literature

Debra Dhillon

Abstract

Recent decades have seen a proliferation of research into the identification and manipulation of question difficulty factors. This paper evaluates three predictive models of question difficulty, each of which provides valuable insights into some of the successes and potential pitfalls involved in the process of delivering an examination question at a specific level of difficulty. In the light of these insights, the viability of developing a unified model of question difficulty applicable across item types and subject domains is assessed. The paper concludes that for any such model to be useful it may have to sacrifice a degree of rule-based precision in favour of flexibility and responsiveness.

Introduction

Imagine the following scenario: you are a 16-year-old student nervously waiting to begin a GCSE Mathematics examination. The invigilator signals that you may begin and you turn over the front page of the paper to discover the following question; '$2^2 + 3^3 = ?$'. Now compare that to the situation where the opening question is as follows;

'Given that, for $n \geq 0$, $I_n = \int_0^1 x^n \, \mathrm{e}^x \, \mathrm{d}x$, prove that

$$I_n = \mathrm{e} - nI_{n-1}, \; n \geq 1.$$'

It does not require a sophisticated imagination to predict that the two questions described above will provoke two very different reactions from candidates and may produce very different levels of performance. Formally understanding precisely why this is so is an issue that strikes at the very heart of the design and implementation of public examinations; that is, what makes an examination question difficult?

Almost anyone who has ever taken a written examination in any subject will believe they intuitively know the difference between an easy question and a difficult one. Similarly, the manipulation of question difficulty is a complex task that many question-setters appear to achieve through application of tacit informal knowledge, or what we might term 'intuition,' and although they are often successful in this endeavour it is nevertheless a somewhat inexact science. If one could formalise these implicit processes it would enable question-setters to target questions at specific levels of difficulty appropriate to particular candidate groups, thereby better assessing the full range of student abilities.

Unfortunately, development of a precise, flexible, and diverse model of question difficulty — one that is of equal utility across item and subject types — is hampered by the fact that the assessment objectives, formats and assessment criteria for examination questions will vary both between and within subject domains. For this reason, much of the research to date has been both subject- and item-specific, developing and testing theories and models to account for differential levels of difficulty within very homogenous sets of item-types. Examples of these are analogy problems and other non-verbal reasoning items such as those prevalent on psychometric tests purporting to gauge general intelligence (Mulholland, Pellegrino & Glaser, 1980; Sternberg, 1977, 1979, 1982), short answer language comprehension questions such as those found on English and Modern Foreign Language papers in the United Kingdom (Pollitt & Ahmed, 1999) and an assortment of problems of the kind that are found on Graduate Record Examinations (GRE) in the United States such as (algebraic) Generating Examples (Katz, Lipps & Trafton, 2002), Sentence Completion (Sheehan & Mislevy, 2001), Analytical Reasoning (Boldt, 1998; Chalifour & Powers, 1988), Verbal Discretes (Adams, Carson & Cureton, 1993), and Reading Comprehension (Freedie & Kostin, 1992).

While some authors have made commendable strides towards developing more generic models of question difficulty (e.g., Pollitt & Ahmed, 1999; Pollitt, Entwistle, Hutchinson & de Luca, 1985), these have been hindered by the fact that any gains in universality have incurred inevitable costs to levels of detail. This partly explains the inadequate pool of published empirical evidence testing the efficacy, practicality, and generality of their application. Acknowledging this is less a criticism than recognition of the inevitable limitations that any generic model must, by definition, contain.

Before exploring these theories in more detail, it is important to define more specifically what is meant by question difficulty. Pollitt *et al*. (1985) and Ahmed and Pollitt (1999), for example, draw several useful distinctions that will be partially adopted in this paper. '*Concept'* difficulty, they argue, is concerned with the inherent conceptual complexity of the subject matter and is determined by the degree to which the concepts involved in a question are abstract or concrete. '*Process'* difficulty concerns the difficulty of the cognitive operations and the degree to which they utilise finite cognitive resources. '*Question'* difficulty refers to the linguistic and structural properties of the question and the appropriate use of mark schemes. Adapting these classifications, the present paper distinguishes between '*intrinsic'* content-bound question difficulty and '*surface'* format-bound question difficulty. Intrinsic difficulty is taken as a more inclusive term, incorporating 'concept' and 'process' difficulty, regarding these as inextricably linked. As such it refers to the difficulty imposed on candidates because of the cognitive demands placed on them, whether conceptual, analytical, memory-based, or otherwise. Furthermore, because the 'concept' difficulty of questions is often highly subject-specific (Ahmed & Pollitt, 1999), determination of intrinsic difficulty factors is also likely to be highly item- or subject-specific. Surface difficulty, which one might expect to be more readily generalised across subjects and item-types, refers to the linguistic, structural and visual style of questions and the flexibility, transparency and reward structure of mark schemes. The contrast between intrinsic and surface difficulty is a particularly important one because many papers that purport to tackle the issue

of question or item difficulty frequently address one of these main parameters without adequately explaining or controlling for the other.

Pollitt *et al*. (1985) further differentiate between legitimate and illegitimate sources of question difficulty. Legitimate sources of question difficulty are those that *intentionally* and *transparently* seek to assess skills or knowledge representative of a level of aptitude or proficiency in a subject. Conversely, illegitimate question difficulty is indicative of a communication failure between two or more of the 'characters' in the assessment dynamic; the question setter, the candidate and the marker. In the latter scenario, a candidate may find a question difficult and hence fail to score marks, not through any lack of proficiency in the subject matter, but because the question fails to convey its specific task demands or makes demands that are beyond the realm of the subject area. While identifying illegitimate sources of difficulty can prove profitable in itself, this paper is primarily concerned with identifying item features that predict legitimate sources of question difficulty.

The present review is intentionally selective, focussing on three of the most insightful models of question difficulty to have emerged from the literature, two of which are item- or subject-specific and one which is more generic, each of which is capable of furthering our understanding of question difficulty.[1] In two main sections it will deal with theories concerning item attributes that range from intrinsic conceptual task demands to syntactic and structural features of question design, focussing in greatest detail on the more theory-driven former. Furthermore, by collating and adapting common themes within this body of work, it seeks to assess the feasibility of developing a unified framework by which question setters can achieve what some already intuitively accomplish; deliver a question at a specific and legitimate level of difficulty. As Pollitt *et al*. (1985) note, this is a task that at present "bears more resemblance to a complex, perceptual art-form than a predictive science" (p. 28).

Intrinsic Question Difficulty

Cognitive Modelling of Answer Generation

It would seem that fundamental to understanding the 'intrinsic' features of an examination question that contribute to its difficulty, is first acquiring an understanding of the inherent cognitive demands that are placed on a candidate in the process of generating an answer. The present section explores the work of several authors who have applied information processing approaches to the deconstruction and subsequent re-construction of certain types of examination questions, ultimately with a view to manipulating item difficulty.

[1] In addition to other largely item-specific predictive models of question difficulty (see, e.g., Adams *et al*., 1993; Attali & Goldschmidt, 1996; Freedie & Kostin, 1992; Sheehan & Mislevy, 2001), which the reader may find of interest, there are a number of related issues concerning item difficulty including research exploring difficulty factors (intrinsic and surface; see, e.g., Fisher-Hoch, Hughes, & Bramley; 1997; Pollitt *et al*., 1985) and theoretical foundations for the distinction between difficult and easy questions (e.g., Malpas & Brown, 1974).

The Cognitive Components Approach

Sternberg (1977, 1979, 1982) has applied a 'cognitive components' approach (Pellegrino & Glaser, 1979) to understanding the series of mental processes involved in solving non-verbal reasoning tasks, particularly analogy problems of the kind illustrated in Figure 1 (see Smith, 1986 for a detailed review of this work). Despite the narrow focus of the model, its instinctive appeal and versatility renders it the derivational basis of some other more generic accounts of question difficulty and merits its early inclusion in this discussion.

Analogy item types, whether verbal, pictorial, or geometric (e.g., Figure 1), take the form "*A is to B as C is to ?*". Sternberg (1977) hypothesised that a series of roughly sequential (and at times cyclical) cognitive steps or 'components' could be identified as integral to the solution of such a problem. A summary of these cognitive components is presented in Table 1. The first step is one of ***encoding***, whereby the candidate translates the simple notation of 'A' and 'B' into the geometric shapes (or words/pictures) that they represent. By a process of ***inference*** the candidate then deciphers the exact nature of the relationship between these two shapes. Subsequently, they recognise, by ***mapping***, the features of C that make it an analogy of A and then deduce the solution to the problem by ***application*** of the inferred A-B relationship on to C. Optionally, they might select and ***justify*** the correct solution from the choice of distracters and finally they must generate a ***response***.

Table 1: Sternberg's (1977) hypothesised cognitive components for the solution of analogy item types

	Cognitive component	Description
❶	**encoding**	*Translating simple notation into representative words or images*
❷	**inference**	*Determining the relationship between A and B*
❸	**mapping**	*Determining the relationship between A and C*
❹	**application**	*Transforming C according to the inferred rule(s)*
❺	**justification**	*Optional stage of checking the answer if it doesn't seem obviously correct*
❻	**response**	*Indicating or presenting the selected answer*

Sternberg (1977, 1979) hypothesised that changes to the complexity of components ❷, ❸ and ❹ (inference, mapping and application) would have the greatest impact on the overall difficulty of geometric analogies. To illustrate this point, let us first examine how this cognitive components model informs the solution process of a comparatively simple analogy problem of the kind illustrated in Figure 1 and compare this to the solution of a more complex analogy problem such as that depicted in Figure 2. One can see that in Figure 1 images A and B may each be described as comprising of two relevant geometric features or 'elements' (Smith, 1986, p. 126): shape and fill. The tasks of perceiving or recognising these might be seen as the cognitive 'sub-components' of Sternberg's encoding (❶)

component. Similarly, one can infer (❷) that the relationship between A and B may be characterised by one or more transformations to some or all of these elements. In this case, image B represents a simple 'reduction' in the size of image A. Mapping (❸) from A to C one observes that C is different but analogous to A. Applying (❹) the single simple transformation of a 'reduction' onto the shape of image C, we can select (❻) distracter 'c' as the correct solution with relative ease.

Figure 1: Example of a simple analogy item type (adapted from Smith, 1986)

Below are two figures (A and B) which go together to make a pair. Then there is a third figure (C) and a choice of five more (a to e). Choose which of the five makes a pair with the third figure, like the first pair.

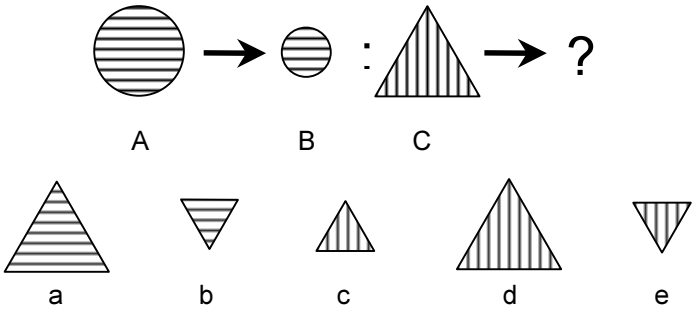
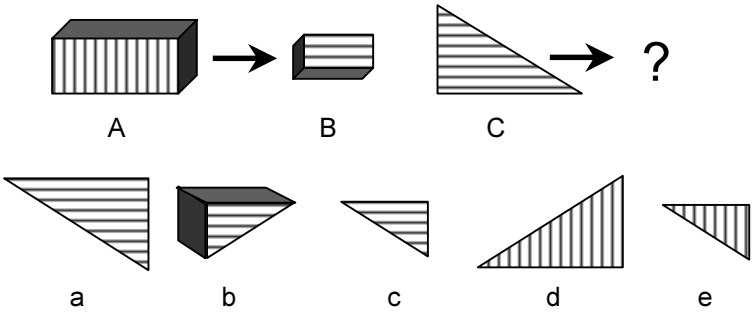


Figure 2 below is an example of a more complex analogy item and illustrates what happens when components ❶ and ❷ are comprised of a number of more complicated sub-components. This time images A and B may be described as comprising of three notable or relevant geometric elements: shape, dimensionality (i.e., 2D or 3D) and fill. Furthermore, inferring the relationship between A and B in this example is more difficult because two out of three of these elements have been subjected to one or more transformation each; the shape has been rotated by 180° and reduced in size, the fill has been rotated by 90°, while dimensionality remains constant. The subsequent components of the task can be similarly dissected into constituent parts for the *mapping* of C, the *application* of the aforementioned transformations to C, and the selection (and possible *justification*) of the *response*. Therefore, not only is the problem more difficult by virtue of the increased number of sub-components involved in the encoding and inference stages, but the student must also do more 'work' for each of the subsequent components of mapping and application. Consequently, selection of the correct answer ('e') becomes more time-consuming and error-prone. In contrast, simply changing the complexity of the images themselves without altering the relationship between them (i.e., manipulating component ❶ (encoding) but not component ❷ (inference)) is unlikely to be nearly as challenging to the student.

Figure 2: Example of a more complex analogy item type

Below are two figures (A and B) which go together to make a pair. Then there is a third figure (C) and a choice of five more (a to e). Choose which of the five makes a pair with the third figure, like the first pair.



Predictive validity

In terms of item design, the advantages of the cognitive components approach are obvious. Not only does Sternberg's (1977) model identify the main sources of differential levels of difficulty between different non-verbal reasoning items, but by identifying the components and sub-components of a question task, one can also attempt to predict the specific level of difficulty of individual questions. Assuming a simple additive model of question difficulty, Sternberg proposed that the difficulty level of a question as a whole was comprised of the sum of the difficulty of each of the individual components (which was in turn dependent upon the difficulty of the sub-components), multiplied by the number of times each component was executed. In other words, one could predict the difficulty of a question by breaking the task down into the smallest of cognitive units, or steps, assigning a uniform difficulty quotient to each, and then simply 'totting' these all up. The greater the number of units, the longer the task would take to complete and the more opportunity one had to make errors along the way.

A systematic and empirical study of analogy problems conducted by Mulholland *et al*. (1980) provides some evidence in favour of this cognitive components approach. However, as well as highlighting the successes of Sternberg's (1977) model of analogy item difficulty, their study also illuminates the model's failures, particularly in its assumption of additive or linear increments in difficulty. Mulholland *et al*. presented participants with various permutations of analogy problems containing between one and three stimulus elements (sub-components of the encoding component), and between one and three transformations (sub-components of the inference component). Subsequently, they measured the participants' solution times and error rates. They found that if the number of transformations (e.g., reduction or rotation) was held constant, increasing the number of elements (e.g., shape and size) produced a *linear additive increase* in solution times. The same was true if the number of elements was held constant and the number of transformations increased. Up to this point, Sternberg's assumption of linearity held true. However, Mulholland *et al*.'s findings on both measures of item difficulty (solution times and error rates) demonstrated an interaction between the number of elements and the number of transformations featured in a problem. Specifically, any increase in solution times

resulting from a greater number of transformations was aggravated by increasing numbers of elements and vice versa. In other words, any increase in item difficulty as a result of greater complexity in either of these components (as reflected in longer response latencies) was compounded by concurrent complexity in the other.

The findings with regard to error rates (as opposed to solution times) were slightly less straightforward. As would be expected from a linear model, higher error rates were incurred with greater numbers of transformations. Interestingly, this was not the case with higher numbers of elements; seeming to confirm, at least in part, Sternberg's predictions that changes to the inference component would have a greater impact on difficulty than changes to the encoding component. However, crucially, it was not merely the number of transformations *per se*, but the number of transformations *per element* that had the greatest impact on item difficulty. Multiple transformations to a single element produced many more errors than did the same number of single transformations to different elements even though the total number of cognitive processes remained constant. Item difficulty on analogy problems, it seemed, could not be predicted by mere numbers of components and sub-components because one type of information processing (e.g., encoding) appeared to interact 'antagonistically' with other types of information processing (e.g., inference).

So why does the assumption of linearity in analogy item-difficulty prediction break down at higher levels of item complexity? Mulholland *et al.*'s (1980) explanation invokes Baddeley's (Hitch & Baddeley, 1976) seminal 'working memory' construct; so pervasive in theories of learning and memory that, as Byrne (2000) notes, some researchers have claimed that it may be the principal operational source of individual differences in cognitive ability. Baddeley proposed that one of the most limiting factors in human information processing is short-term memory capacity, more specifically the capacity of the 'working' sub-system of short-term memory in which new information is briefly held and rapidly processed in conjunction with learned information retrieved from long-term memory. It is well-documented that the average numbers of 'bits' of information that can be stored in working memory is seven (plus or minus two) at any given time, although individuals will differ on exact numbers (Miller, 1956; Baddeley, 1994). Should one encounter another 'bit' of information that requires processing during this time, the efficiency of the whole system deteriorates because older information becomes more difficult to retrieve and new information becomes subject to concomitant loss or 'decay'. Consequently, any task involving working memory (which would appear to be most cognitive tasks) will take longer to complete and be more prone to errors when working memory capacity is saturated. Analogy items comprising of multiple transformations to single elements are more difficult than those comprising of single transformations to multiple elements because in the former the transformations must be performed concurrently whereas in the latter they may be performed sequentially. The first type of processing requires more information to be stored in working memory simultaneously.

Further support for the role of working memory in the determination of item difficulty has come from investigations into a variety of other tasks such as letter series completion (Kotovsky & Simon, 1973), numerical analogies and number series completion (Holzman, Pellegrino & Glaser, 1982, 1983).

Furthermore, working memory appears to be a particularly limiting factor in children's cognitive performance, many of whom have frequent experience of the public examination system. As such, one could argue that in terms of question difficulty, working memory appears to be both the strongest and weakest link in the chain of information processing. Appreciation of its role is central to the validity of any predictive model of difficulty.

The instructive simplicity of Sternberg's (1977) model of item difficulty belies another important shortcoming. By neatly dissecting a problem into cognitive 'chunks' or sub-components, whether broad or specific, one fails to adequately consider the differences that will occur in the complexity of individual sub-components by virtue of the type of task they may involve. For example, the model assumes that the cognitive demand involved in reflecting an image is equivalent to that involved in enlarging it. This is a point that Whitely and Schneider (1981) recognised as crucial to any model of analogy item difficulty. They found that one could better predict difficulty measures of geometric analogy items if, instead of merely counting the numbers of transformations, one distinguished between different types of transformations, particularly spatial versus non-spatial ones, and weighted them accordingly.

Concluding comments

It is clear that Sternberg's (1977) cognitive components approach to modelling the difficulty of non-verbal reasoning items is not without its limitations. It does not adequately compensate for the limitations of working memory (Holzman *et al*., 1982, 1983; Mulholland *et al*., 1980) and it treats all of the sub-components within a component (i.e., the transformations within the inference component) equally when their impact on item difficulty is plainly not uniform. Neither of these problems is necessarily insurmountable and both might be remedied with some adjustment although the exact form of the required modifications and their ease of implementation are questionable. Therefore, while Sternberg's cognitive components model is no panacea to the tricky issue of predicting question difficulty, it does enable one to derive a clear operational guide to the construction of analogy problems and many other non-verbal reasoning items, albeit with some fine-tuning. The next section examines whether a similar approach might be fruitfully extended to other question types.

Pollitt and Ahmed's (1999) Model of the Question Answering Process

Over recent years, a number of researchers from the same research group have endeavoured to extend and adapt the item-specific cognitive components approach to broader models with more generalised applicability. The work of Alastair Pollitt has featured heavily in this arena (Ahmed & Pollitt, 1999; Pollitt & Ahmed, 1999; Pollitt *et al*., 1985; Pollitt & Hutchinson, 1987). The Language Comprehension Model (Pollitt & Ahmed, 1999) is just such an attempt at diversification which seeks to explain the sources of difficulty observed in reading comprehension questions with a view to representing the key psychological processes involved in generating answers to written questions in general. Note that while the model draws heavily from the component view of mental processes

advocated by Sternberg (1977), in contrast to Sternberg's model which relates exclusively to non-verbal reasoning items, it deals more generally with written question types that for the most part require verbal responses, although it does not explicitly restrict itself to only these.

At its simplest Pollitt and Ahmed's (1999) model defines the sequence of conscious and unconscious mental activities that the student must engage in when responding to a reading comprehension item. Such items typically comprise a piece of text followed by short-answer questions. After reading the text in question the student proceeds by *understanding* the question, *searching* their mental representation of the read text for relevant part(s), *interpreting* these mental representations, and finally *composing* the answer. Although these stages are then further broken down into as many as 44 smaller steps, which are no doubt more subject specific, the authors propose that the basic processes are essentially the same irrespective of subject matter (Ahmed & Pollitt, 1999; Pollitt & Ahmed, 1999). Pollitt and Ahmed argue that the secret to translating this model from language comprehension to all examination questions is simply in replacing 'text' with 'subject' and making a few logical amendments. More specifically, they contend that understanding a subject is simply a macro-level version of comprehending a language or, as they state, "understanding a story" (p. 5). Consequently, it may be similarly broken down into corresponding processes.

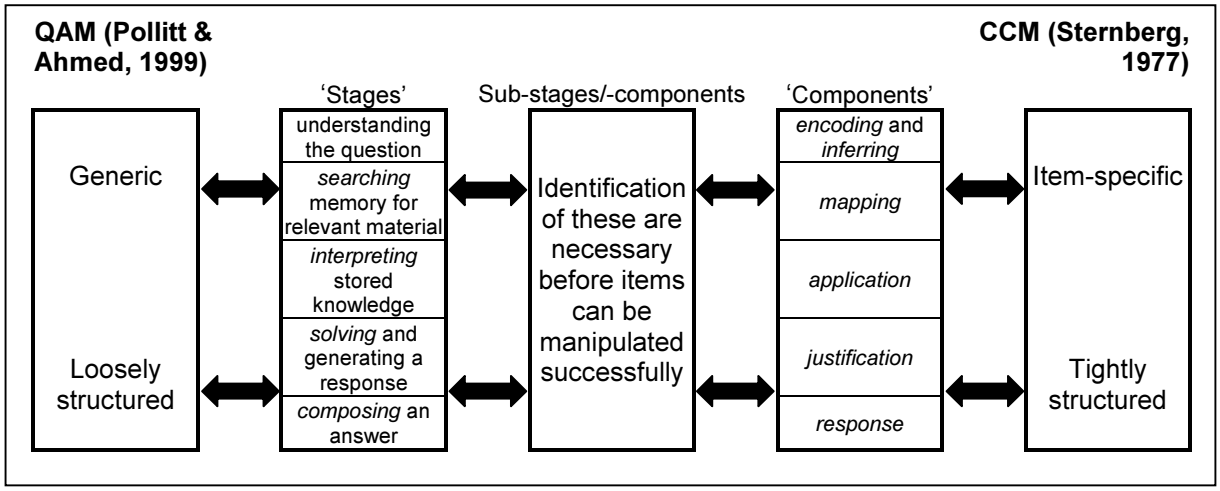
Table 2 shows how these supposedly equivalent processes are conceptualised: after their prior ***learning*** of the subject the student proceeds by ***reading*** and understanding the question, ***searching*** their knowledge of the subject for relevant topics, strategies or skills, ***interpreting*** this material, and finally the process of 'composing an answer' is sub-divided into generating a ***solution*** and ***composing*** a written answer. Depending on the precise task demands of the question, these processes may be conscious or unconscious, they will often occur in rapid succession, may sometimes repeat cyclically, and frequently they may subsume or overlap with each other.

Table 2: Development of the Question Answering Model ('QAM') from the Language Comprehension Model

Language Comprehension Processes	General Question Answering Processes or 'Stages'	'QAM' Stage Labels
reading the text	learning the subject	*Learning*
understanding the question	understanding the question	*Reading*
searching the mental representations of the text for the relevant part(s)	searching (and accessing) relevant aspects of the subject from memory	*Searching*
interpreting the parts of the mental representation	re-interpreting stored knowledge to match the question	*Interpreting*
composing the answer	generating a response to the question	*Solving*
	writing an answer for the examiner to read	*Composing*

When Pollitt and Ahmed's (1999) model of the question answering process is placed side-by-side with Sternberg's (1977) cognitive component analysis of analogy problem solving model we are faced with an interesting paradox, as depicted in Figure 3. While both models share inescapable similarities, one might also perceive them as lying on diametrically opposite extremes of at least two, roughly parallel continua: the cognitive components model is intentionally subject- and item-specific while the question answering process is intentionally generic, transferring across subject and item types. The cognitive components model is tightly structured and prescriptive at a micro-level, while in relative terms the question answering model is loosely structured and vague at a macro-level. There are nonetheless common features to both models.

Figure 3: Comparison of Pollitt and Ahmed's Question Answering Model ('QAM'; 1999) with Sternberg's Cognitive Components Model ('CCM'; 1977)



Pollitt and Ahmed (1999) highlight the parallels between Sternberg's cognitive components and their generalised question answering stages. The comparisons identify the functional similarities between the two approaches well. It is apparent, for example, that the process of 'understanding the question' seems to engender similar mental activities to Sternberg's *encoding* and *inference* components. Furthermore, one cannot escape the fact that Sternberg's model works best when one identifies the sub-components of the question task as well as the components. Item difficulty will then depend upon the number and interaction of different types of sub-components. Similarly, Ahmed and Pollitt (1999) recognise that stage one of their question answering model ('understanding the question') consists of numerous mental operations that contribute to the cognitive demands and conceptual complexity of the question. These they liken to 'hurdles' that the student must overcome in order to produce an answer, the 'number' and 'height' of which will contribute to question difficulty. Such a comparison also exposes an important weakness of the question answering model: we have previously seen that minor specific adjustments to both the encoding and the inferring components, especially in combination, can yield quite large increments in difficulty on analogy problems. Subsuming these components into one process, which the question answering model does, is unlikely to be helpful when it comes to determining intrinsic cognitive sources of question difficulty.

Predictive validity

Pollitt and Ahmed's (1999) model of the question answering process seems robust in so far as it seeks to draw on a variety of theoretically relevant fields within cognitive psychology concerned with learning and memory, problem solving, attention and language comprehension. It is both complimented and driven by empirical, though largely *post hoc*, investigations into Sources of Difficulty and Easiness (SODs and SOEs) in a selection of subjects (Ahmed & Pollitt, 1999; Hughes, Pollitt & Ahmed, 1998; Pollitt *et al*., 1985). In addition, awareness of the processing stages advised by Pollitt and Ahmed's (1999) model will not only facilitate the identification (and manipulation) of legitimate sources of difficulty but also the identification (and elimination) of illegitimate sources of difficulty. The question-setter can then better foresee hurdles at which the student will fall through no fault of their own. Nevertheless, the utility of the model is not empirically confirmed and may be a major limiting factor, especially for questions that are answered in a less verbal manner. This raises the question of how conveniently the generic Question Answering Model can be applied to the manipulation of difficulty of specific novel item types.

Concluding comments

Where Pollitt and Ahmed's (1999) model seems undoubtedly useful is in signalling the more generic sub-processes within the whole question answering process that item designers should keep in mind when attempting to formulate a question to a certain level of difficulty. In other words, the major advantage of the model is that unlike Sternberg's (1977) highly item-specific model, it provides an invaluable starting point to deconstructing the answering process of almost any kind of question; a pre-requisite to any attempted reconstruction of versions of varying difficulty. If the question answering model truly is a generic path-way describing the cognitive behaviour of question-takers, then it should facilitate this task.

It is at this point, however, that one is struck by its generality. Application of the model as it is laid out above, without any task- or subject-specific elaboration does not seem to be particularly beneficial because without identification of the idiosyncratic sub-stages, any combination of which might influence demand differently, the process appears unhelpful. Unfortunately, identification of these is not always immediately obvious. They must often be thought through oneself by actually proceeding to solve the problem. The reader might have noticed that if they have not already paused to tackle some of the questions depicted in the paper so far, the compulsion to do so is difficult to resist because only in doing so do the components and sub-components of these problems become more tangible. The danger is, of course, that these components and sub-components may differ between individuals depending upon the solution strategy adopted.

Cognitive Modelling of Algebraic Word Problems

A large and varied body of research has come about as a result of academic interest in various aspects of item types found on the Graduate Record Examinations (GRE) in the United States. Within

several studies to emerge from the same research team, one in particular has sought to systematically identify and manipulate features of the Generating Examples item type that may predict question difficulty (Katz *et al*., 2002). While this work is admittedly item-specific, it demands inclusion because of its rigorous empirical foundations and potentially far-reaching implications.

The Generating Examples (GE) item type represents a class of mathematical (algebraic) word problems in which examinees must generate examples of potential solutions that simultaneously meet various mutual constraints. Figure 4 below provides a typical example of just such a problem in which implicit constraints are imposed on any (correct) solution although it is apparent that some constraints, such as the requirement for the numbers of stamps to be positive integers, are more implicit than are others.

Figure 4: A Generating Examples (GE) item (reproduced from Katz *et al*., 2002, p. 2)

> Jamal wants to buy some standard-issue stamps at $0.25 each and some commemorative stamps at $0.40 each. He wants more than 25 stamps, including at least two of each, but can spend no more than $10.00
>
> **Question: What is one possibility for the number of standard-issue stamps and commemorative stamps that Jamal could buy?**

The key to answering this question correctly is in explicitly recognising the implicit constraints contained within the wording of the problem; equivalent to *encoding* and *inferring* in the cognitive components paradigm or to *reading* (and understanding) in the question answering model. One way that the problem solver may do this is by representing these constraints in the form of algebraic equations or inequalities such as those depicted in Figure 5. However, according to Katz *et al*. (2002) the most notable feature of the answering process is that despite the algebraic nature of the problem, one cannot derive a unique correct answer by straightforward application of a *standard* algebraic technique such as the solution of simultaneous equations. Essentially, the problem posed in Figure 4 is 'under-determined' because at least one of the constraints (such as the fact that the total number of stamps must exceed 25) is not an 'equation' at all, but an inequality. Katz *et al*. argue that in order to answer such a problem correctly one must generate examples of potential solutions and then check that these do not violate any of the constraints. One might call this a 'bottom-up' method as it involves generating a possible solution and then working backwards. If the problem solver's first estimate is not correct — and no matter how sophisticated their algebra skills, the problem does not provide enough information to guarantee that it will be — they will have to try again.

Figure 5: Explicit algebraic representation of constraints posed in Figure 4

$$S \text{ and } C \text{ are integers} > 1$$
$$S + C > 25$$
$$0.25S + 0.40C \leq 10.00$$

The informal and ostensibly crude 'bottom-up' cognitive model described by Katz *et al*. (2002) for solving GE items has been variously termed a 'generate-and-test' strategy (Nhouyvanisvong, Katz, & Singley, 1997) and a 'guess-and-check' heuristic. Yet its application to the solution of GE items, precisely because of their under-determined nature, has shown to be both effective (Nhouyvanisvong *et al*., 1997) and widespread (Katz & Berger, 1995). This is even the case among mathematically able students who are entirely capable of implementing more traditional and skilful algebraic manipulations (Katz, Bennett, & Berger, 2000; Tabachneck, Koedinger, & Nathan, 1995).

By cognitively modelling the stages of the generate-and-test method outlined above, Katz *et al*. (2002) proposed that one should be able to increase the difficulty of a GE item, quite simply, by increasing the amount of testing the student must do. Both the cognitive load and the potential for errors would consequently be increased. In practice, one might achieve this by manipulating one or both of two processes. First, one could increase the number of constraints against which any estimate must be checked. Second, reducing the solution density of the item (the ratio of correct solutions to all possible answers) and hence also reducing the likelihood of a student generating a correct example, would require the student to generate further solutions for testing.

At first sight, it would seem reasonable to suppose that increasing the number of constraints embedded within the item stem would serve both of these purposes, requiring the student to test their estimate against yet another constraint, and further limiting the range of responses that will correctly satisfy all conditions. Yet as these and other authors note (Bennett, Morley, Quardt, Rock, & Katz, 1999; Nhouyvanisvong & Katz, 1998), not all constraints serve the same purpose. Some constraints, known as *generator constraints*, will facilitate the process of generating a potential solution without increasing the testing load and will consequently have little effect on item difficulty, indeed sometimes making an item easier. For example, addition of the constraint that **C > 14** to the problem in Figure 4 (Katz *et al*., 2002), is likely to focus the initial estimate of commemorative stamps to a number greater than 14 and make testing of the estimate against it superfluous. Other constraints, which Nhouyvanisvong and Katz identify as *verifier constraints*, are used to verify the accuracy of a solution. By increasing the amount of testing required, they serve to increase cognitive load and the potential for mathematical errors. The constraint that **S > 10C**, for example, again applied to the problem in Figure 4, does not help in the generation of a potential solution, but does increase the chances of the student making a mathematical error when they come to checking their estimate(s).

The solution density of a problem may be reduced either by addition of a constraint (generator or verifier) or by changing an existing constraint. Using the previous examples, one can see that addition of the generator constraint that **C > 14** reduces the solution density of the problem by making all estimates where **C ≤ 14** incorrect. Addition of the verifier constraint that **S > 10C** or changing the constraint that **S, C > 1** to **S, C > 2** should have a similar effect. Clearly, reducing the number of correct solutions, whichever method one employs, should make a GE item more difficult. As will be demonstrated in the next section, this issue is not quite this simple, but suffice it to say that the basic premise of this prediction is as described.

In summary, Katz *et al*. (2002) predict that the addition of generator constraints to the item stem of GE item types should have little or no effect on item difficulty. In contrast, the addition of verifier constraints should increase item difficulty by adding to the amount of testing that is required for each potential solution that is generated. If either constraint also reduces the solution density of the problem, thereby reducing the chances that the estimate is correct, the difficulty of the item will be further heightened.

Predictive Validity

At least two studies have provided *post hoc* evidence of differential functioning of verifier and generator constraints such that the former were found to be associated with GE item difficulty while the latter were not (Bennett *et al*., 1999; Nhouyvanisvong & Katz, 1998). However, because neither study systematically manipulated GE items according to the number and type of constraints embedded within the item stem, they provide only weak evidence of a cause-and-effect relationship.

Katz *et al*. (2002), on the other hand, did precisely this. After presenting participants with a pool of GE items that varied independently according to the number of verifier constraints in the stem and/or solution density they found that more verifier constraints and lower solution densities led to more difficult and time-consuming items. However, the predictive success of solution density was tempered by the method used to manipulate it. Item difficulty was not affected if the change in solution density was due to the addition of a generator constraint. To understand why this is the case, let us re-examine the effect of adding the generator constraint that **C > 14** to the problem that was depicted in Figure 4. In theory, this should make the problem more difficult by making all estimates where **C ≤ 14** incorrect, thereby reducing solution density. In practice, however, the constraint diverts the student away from making these incorrect initial estimates — that is, it limits the range of considered solutions. Thus the student modifies their estimating behaviour, consciously or unconsciously, to compensate for any change in solution density of the problem, making them no more likely to actually offer an incorrect solution or to consequently need to generate another solution for testing. This is an important point and illustrates the futility of developing hard and fast rules for the manipulation of item difficulty in the absence of any real understanding of how the subsequent changes might influence students' cognitive processing or test-taking behaviours.

Concluding comments

It is not a controversial argument to suggest that increasing the amount of 'work' that a student must do in order to answer a question will make that question more difficult. The cognitive modelling of the commonly used generate-and-test strategy for solving GE item types has enabled Katz *et al*. (2002) to identify which out of two ostensibly similar additional features of the item stem actually result in increased cognitive load; generator constraints or verifier constraints. Similarly, by noting the simultaneous and often mutually mitigating effects that certain constraints have, both on solution density and on solution generation behaviour, one can avoid making overly simplistic predictions regarding item difficulty. In other words, cognitive modelling of this kind highlights the critical distinction between function and form.

Surface Difficulty

Surface difficulty is format-bound and refers to the linguistic, structural and visual features of questions and the flexibility, transparency and reward structure of mark schemes, both of which may either temper or exacerbate the intrinsic difficulty of a question (Fisher-Hoch *et al*., 1997; Pollitt *et al*., 1985). Such factors might be seen to encompass both the 'scaffolding' provided to students to support the answering process and features of the mark scheme that legitimately reward the student where appropriate, while allowing for consistency (and reliability) in marking.

Surface-Level 'Scaffolding' Features

Empirical work based on Pollitt *et al*.'s (1985) extensive research has lead to the identification of surface-level Sources of Difficulty (SODs) and Sources of Easiness (SOEs) in examination questions for 16-year-olds in a number of subjects. Some of the subject-specific SODs and SOEs derived from these post-hoc analyses of performance data have been confirmed by experimental studies which have attempted to manipulate questions for trial with children (e.g., Ahmed & Pollitt, 1999; Fisher-Hoch & Hughes, 1996) although their relative contributions to item difficulty vary. These authors have further attempted to identify general SODs and SOEs that they argue can be applied across almost all subjects. Some of these are represented in Table 3 below.

It is these generalised surface difficulty factors that remain relatively untested. Moreover, while there is a logical appeal to many of them, we have already observed that complex cognitive processes are rarely amenable to simple transcribed manipulations and when they are they can sometimes be vulnerable to surprise developments. In the absence of any serious theoretical foundations, these as yet untried and untested surface features are liable to have similarly unexpected interactions with intrinsic difficulty manipulations. At present, however, they represent the best available information and as such are presented in this paper for the sake of completeness.

Table 3: General Surface-Level Sources of Difficulty (SODs) and Sources of Easiness (SOEs), adapted from Ahmed & Pollitt (1999) and Fisher-Hoch *et al*. (1997)

SODs and SOEs	Description
Distractors in wording of question	Information appears in question that was not required and may distract from relevant information.
Highlighting	Key words and phrases, or of command words that tell the candidate what to do.
Density of presentation	Too much information condensed into too small an area.
Technical terms	Use of technical terms where everyday language would suffice.
Response prompts	Structure, organisation or content of answer is prompted in question or in cues.
Paper layout/sequence of questions	Physical organisation of the question ordering and/or numbering could support or hinder candidates.
Paper space for response	Style and size of space allocated for response not appropriate.
Ambiguous resources	Unclear resources (diagram, graph, table etc.) affecting performance.
Provision of leaders, cues, clues, or examples	Question cues candidate into particular data or strategy.
Bulleting of information	Key constraints, ideas, or requirements bulleted or similarly clarified

Mark Schemes

Devising an appropriate mark scheme to complement each question on an examination paper and of the paper as a whole is as important in determining student performance (and hence item difficulty) as constructing the questions is. Unless the intrinsic demands of a question are reflected in the reward structure of the mark scheme even a successful manipulation to intrinsic item difficulty will be a hollow victory because candidates' marks will not be a valid reflection of their performances.

Unfortunately, controlling the legitimacy, flexibility, transparency and reliability of the mark scheme can be highly problematic because some of these requirements may at times be in opposition to each other. A recurrent theme within the question difficulty literature relates to the important constructs of *problem space* (e.g., Pollitt & Ahmed, 1999) and *outcome* or *solution space* (e.g., Katz *et al*., 2002; Marton & Saljo, 1976; Pollitt & Ahmed, 1999). Problem space refers to the range of relevant knowledge, skills, strategies, memories, concepts, relationships, and processes, any combination of which might lead to a solution to the problem at hand. Marton & Saljo's (1976) important parallel construct of solution space signifies the range of possible answers that might be offered by students in response to a problem.

Crucially, it is only by predicting the full scope of problem spaces — all combinations of the knowledge and strategies that might legitimately be applied to solving the problem — that we can predict all of the possible solutions that students may give such that, "a perfect mark scheme is an evaluative description of the whole outcome space" (Pollitt & Ahmed, 1999, p. 11). In practise there is no doubt that achieving this can be extremely challenging, but one can aim to optimise a mark scheme by accommodating, as far as possible, both the problem and outcome spaces for each question. One might reasonably predict that the success of this endeavour will be determined in part by the size of the problem and outcome spaces inherent to the particular problem and the degree of expertise the

question setter has of the subject matter being assessed. Perhaps the most important factor, however, is the degree to which the 'expert' who sets the question can anticipate the methods of the relative 'novices' who attempt it.

At present, for the preparation of question papers in the AQA, attempts to bridge the expert-novice divide are subsumed by the role of the Scrutineer (formerly Assessor). This involves "checking that the questions can be answered in the time allowed, that there are no errors in the question paper and that the mark scheme is in line with the question paper. In most instances *the paper will need to be worked through from a candidate's perspective*" (AQA Question Paper Preparation Procedure File, 2002, p. 20, emphasis added).

Summary and Conclusions

Towards a Unified Framework of Question Difficulty?

Pollitt & Ahmed's (1999; Ahmed & Pollitt, 1999; Hughes *et al*., 1998) work in further modifying Sternberg's ideas has yielded profitable contributions to understanding the generalised item parameters that dictate difficulty, although at a cost to specificity. It does not take an enormous leap of the imagination to realise that if simply failing to differentiate between the various specific transformations inherent to an analogy problem can lead to miscalculations of item difficulty, then generalising across broad subject areas and topics will further compound this problem. As a case in point, failing to differentiate between, (i) the mental activities involved in comprehending a piece of text and composing an answer, and (ii) those involved in comprehending (and executing) the solution to a simultaneous equation, will vastly oversimplify a necessarily task-specific process. This issue is a practical observation that is more evident if one tries to globalise Pollitt and Ahmed's model to more heterogeneous problem solving tasks. It is entirely possible that the question answering model is a more useful tool than the cognitive components approach in questions requiring more verbally constructed responses, which would imply that any dichotomy between the models is more one of item-types (verbal versus non-verbal) than degrees of specificity. Alternatively, it may be that it is only ever intended as a preliminary guide and that it is necessarily with the aid of both qualitative protocol analysis and quantitative subject- or item-level analysis, that it can be explicitly and usefully employed to manipulate specific item-types.

Similarly, by virtue of his detailed piece-meal operational analysis of answer generation, Sternberg (1977, 1979) has shown that a cognitive components approach to modelling item difficulty is a fruitful means of constructing questions to prescribed levels of difficulty. Nevertheless, the very strength of his model — its specificity — also appears to be its weakness because in order to extend the model to more heterogeneous problem solving tasks one must conduct a similarly painstaking task-specific analysis for each one. Whichever model one employs, therefore, it is likely that the item under manipulation and the subject area it assesses will be the key factors determining utility. So what this

discussion boils down to is this: in order for the adjustment or calibration of intrinsic concept-bound question difficulty to be successful, for each item type there will inevitably require careful step-by-step scrutiny of intrinsic mental operations guided by one or both of the models discussed above. Furthermore, this is a process which will often require the input of detailed specialist subject knowledge, especially for items that are peculiar to a subject domain.

The models discussed in this paper also present other challenges to a unified framework of item difficulty and demonstrate that the difficulty of a question is more than merely the sum of its parts. One will recall, for example, Mulholland *et al*.'s (1980) careful and systematic test of Sternberg's (1977) predictive model of analogy item difficulty and the circumstances under which the assumptions of linearity were violated. Specifically, the major limiting factor in students' performances was found to be working memory. Item difficulty was not only predicted by the number and complexity of the cognitive components required for solution, but also by the way those component tasks were combined and the simultaneously ensuing demands they placed on candidates' finite cognitive resources.

Sternberg's (1977) predictions of item difficulty could be similarly complicated because the model failed, for example, to take into account the difference between spatial and non-spatial transformations (Whitely and Schneider, 1981), a criticism that Pollitt and Ahmed's model seems particularly susceptible to. On the other hand, there are other, less obvious advantages to their model that can be easily overlooked. For example, it recognises that some questions will be more difficult than others not because of the minute 'mechanics' of the cognitive tasks, but because before the task can even be attempted the student must understand what is required of them and then search for or construct an answering strategy. They must understand the demands of the question and define its problem space; determine the 'how' and 'what' of the task before they can tackle the actual 'work' — a point echoed within the vast problem-solving literature (e.g., Pólya, 1945). Questions that specifically test how well students can identify or construct novel or unfamiliar problem-solving strategies, in addition to the knowledge required to actually apply the strategies, are likely to be less sensitive to 'nuts and bolts' manipulations to cognitive components if these only influence the latter.

As a purely speculative point, changing the difficulty of such problem-solving questions may instead be reliant on changing the size or familiarity of the problem space so that students can more or less easily land on relevant material within the mass of irrelevant subject matter. Perhaps one might describe this as changing the 'problem density' of a problem in a manner akin to, but undoubtedly as complicated as Katz *et al*.'s (2002) manipulations to solution density. Thus, a question with a low problem density might require the student to search a large body of learned material to find a unique or relatively unfamiliar strategy, skill, or piece of information and may be difficult even if the ease of the actual application of that material remains simple. One does not pretend that any relationship between problem density and question difficulty is straightforward, but it may nevertheless be an issue worthy of empirical investigation.

Finally, one should not overlook the contributions that authors such as Katz *et al*. (2002) have made to the understanding of question difficulty. Although their generate-and-test model may have at times seemed technically convoluted, with little apparent application to generic issues of question difficulty, their prediction that the solution density of a GE problem should impact on item difficulty may have important wider implications. It is possible that solution density influences the difficulty of not only Generating Examples items, but the difficulty of all questions where candidates may respond by estimating or guessing. If so, anyone attempting to estimate the solution density of a problem is likely to encounter similar obstacles as Katz *et al*., who defined solution density as the ratio of (finite) correct answers to all possible (infinite) answers. In practice, of course, they limited the solution space to all possible *reasonable* answers, but definitions of reasonable may vary considerably from examiner to examiner and, more importantly, between examiner and candidate.

In conclusion, developing a simple, rule-dependent model of question difficulty that can be precisely transferred between item types may be an unrealistic goal. As we have seen, cognitive processes are interactive and sometimes antagonistic and may not be amenable to simplistic deconstruction. Certainly, as discussed, some lessons can and should be learned from the models already developed. It may, for example, be possible to identify common themes and pitfalls, to develop a more fluid and dynamic model of question difficulty which does not claim exacting precision or hard and fast rules but is adaptive to the peculiar cognitive demands of individual questions and subject domains. Working with such a model may require specialist knowledge of question demands and answering processes from 'experts' with the insight to foresee the responses of 'novices'; no easy task. However, giving straightforward and effective advice to test-setters without such interactive input will be difficult. Lastly, development of any such model will require empirical support and validation in the form of experimental studies that systematically manipulate questions for subsequent testing with student populations. Those seeking short cuts may be sadly disappointed.

Debra Dhillon
February 2003

References

Adams, R., Carson, J., & Cureton, K. (1993). *Item Difficulty Adjustment Study: GRE Verbal Discretes* (GRE Board Professional Report No. 89-04P). Princeton, NJ: Educational Testing Service.

Ahmed, A. & Pollitt, A. (1999). *Curriculum Demands and Question Difficulty*. Paper presented at IAEA Conference, Slovenia, May.

Attali, Y. & Goldschmidt, C. (1996). The Effects of Component Variables on Performance in Graph Comprehension Tests. *Journal of Educational Measurement*, *33*(1), 93-105.

Baddeley, A. (1994). The magical number seven: Still magic after all these years? *Psychological Review*, *101*(2), 353-356.

Bennett, R. E., Morley, M., Quardt, D., Rock, D. A., & Katz, I. R. (1999). *Evaluating an underdetermined response type for the computerized SAT* (ETS Research Report No. 99-22). Princeton, NJ: Educational Testing Service.

Boldt, R. F. (1998). *GRE Analytical Reasoning Item Statistics Prediction Study* (GRE Board Professional Report No. 97-18P). Princeton, NJ: Educational Testing Service.

Byrne, M. D. (2002). A Computational Theory of Working Memory. Available online at *http://www.acm.org/sigchi/chi96/proceedings/doctoral/Byrne/mdb_txt.htm*

Chalifour, C. & Powers, D. E. (1998). Content Characteristics of GRE Analytical Reasoning Items (GRE Board Professional Report No. 84-14P). Princeton, NJ: Educational Testing Service.

Fisher-Hoch, H. & Hughes, S. (1996). *What Makes Mathematics Exam Questions Difficult?* Paper presented at the British Educational Research Association Annual Conference, Lancaster, September.

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, York, September.

Freedie, R. & Kostin, I. (1992). *The Prediction of GRE Reading Comprehension Item Difficulty for Expository Prose Passages for each of Three Item Types: Main Ideas, Inferences and Explicit Statements* (GRE Board Professional Report No. 87-10P). Princeton, NJ: Educational Testing Service.

Hitch, G. J. & Baddeley, A. D. (1976). Verbal reasoning and working memory. *Quarterly Journal of Experimental Psychology*, *28*(4), 603-621.

Holzman, T. G., Pellegrino, J. W., & Glaser, R. (1982). Cognitive dimensions of numerical rule induction. *Journal of Educational Psychology*, *74*, 360-373.

Holzman, T. G., Pellegrino, J. W., & Glaser, R. (1983). Cognitive variables in series completion. *Journal of Educational Psychology*, *75*, 603-618.

Hughes, S., Pollitt, A., & Ahmed, A. (1998). *The development of a tool for gauging the demands of GCSE and A Level exam questions*. Paper presented at the British Educational Research Association Annual Conference, Belfast, August.

Katz, I. R., Bennett, R. E., & Berger, A. (2000). Effects of response format on difficulty of SAT mathematics items: It's not the strategy. *Journal of Educational Measurement*, *37*(1), 39-57.

Katz, I. R. & Berger, A. (1995). *Strategies underlying score differences on SAT mathematical items*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA., April.

Katz, I. R., Lipps, A. W., & Trafton, J. G. (2002). *Factors Affecting Difficulty in the Generating Examples Item Type* (GRE Board Professional Report No. 97-18P). Princeton, NJ: Educational Testing Service.

Kotovsky, K. & Simon, H. A. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology*, *4*, 399-424.

Malpas, A. J. & Brown, M. (1974). Cognitive demand and difficulty of GCE O-level mathematics pretest items. *British Journal of Educational Psychology*, *44*(2), 155-162.

Marton, F. & Saljo, R. (1976). On qualitative differences in learning: 1 – Outcome and Process. *British Journal of Educational Psychology*, *46*, 4-11.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, *12*, 252-284.

Nhouyvanisvong, A. & Katz, I. R. (1998). The structure of generate-and-test in algebra problem solving. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Nhouyvanisvong, A. & Katz, I. R., & Singley, M. K. (1997). *Toward a unified model of problem solving in well-determined and underdetermined algebra word problems*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, March.

Pellegrino, J. W. & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. In R. J. Sternberg and D. K. Detterman (Eds.), Human Intelligence: Perspectives on its Theory and Measurement. Norwood, NJ: Ablex Publishing Corporation.

Pollitt, A. & Ahmed, A. (1999) A New Model of the Question Answering Process. Paper presented at the International Association for Educational Assessment, Bled, May.

Pollitt, A., Entwistle, N., Hutchinson, C., & de Luca, C. (1985). *What Makes Exam Questions Difficult?* Edinburgh: Scottish Academic Press.

Pollitt, A. & Hutchinson, C. (1987). The validity of reading comprehension tests: What makes questions difficult. In D. Vincent, A. K. Pugh and G. Brooks (Eds.), *Assessing Reading*. Basingstoke: Macmillan Education.

Pólya, G. (1945). *How to solve it.* London: Penguin Books.

Sheehan, K. M. & Mislevy, R. J. (2001). An Inquiry into the Nature of the Sentence-completion Task: Implications for Item Generation (GRE Board Professional Report No. 95-17bP). Princeton, NJ: Educational Testing Service.

Smith, P. (1986). Application of the information processing approach to the design of a non-verbal reasoning test. *British Journal of Educational Psychology*, *56*, 119-137.

Sternberg, R. J. (1977). *Intelligence, Information Processing, and Analogical Reasoning: The Componential Analysis of Human Abilities*. Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1979). The nature of mental abilities. American Psychologist, 34, 214-230.

Sternberg, R. J. (1982). Reasoning, problem solving and intelligence. In R. J. Sternberg (Ed.), *Handbook of Human Intelligence*. New York: Cambridge University Press.

Tabachneck, H. J. M., Koedinger, K. R., & Nathan, M. J. (1995). A cognitive analysis of the task demands of early algebra. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Whitely, S. E. & Schneider, L. M. (1981). Information structure for geometric analogies: a test theory approach. *Applied Psychological Measurement*, *5*, 383-397.