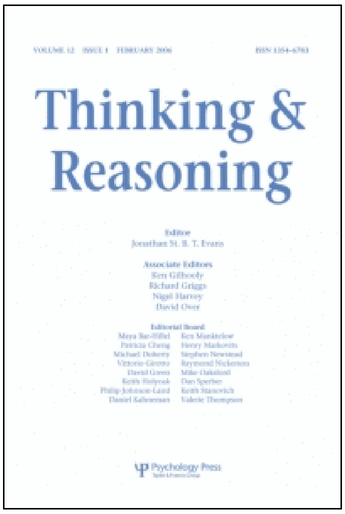

Thinking & Reasoning
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713685607

Predicting the difficulty of complex logical reasoning problems

Stephen E. Newstead; Peter Bradon[a]; Simon J. Handley[a]; Ian Dennis[a]; Jonathan St. B. T. Evans[a]
[a] University of Plymouth, UK

To cite this Article Newstead, Stephen E. , Bradon, Peter , Handley, Simon J. , Dennis, Ian and Evans, Jonathan St. B. T.(2006) 'Predicting the difficulty of complex logical reasoning problems', Thinking & Reasoning, 12: 1, 62 — 90

To link to this Article: DOI: 10.1080/13546780542000023

URL: http://dx.doi.org/10.1080/13546780542000023

Psychology Press
Taylor & Francis Group

Predicting the difficulty of complex logical reasoning problems

Stephen E. Newstead, Peter Bradon, Simon J. Handley,
Ian Dennis, and Jonathan St. B. T. Evans
University of Plymouth, UK

The aim of the present research was to develop a difficulty model for logical reasoning problems involving complex ordered arrays used in the Graduate Record Examination. The approach used involved breaking down the problems into their basic cognitive elements such as the complexity of the rules used, the number of mental models required to represent the problem, and question type. Weightings for these different elements were derived from two experimental studies and from the reasoning literature. Based on these weights, difficulty models were developed which were then tested against new data. The models had excellent predictive validity and showed the relative influence of rule based factors and factors relating to the number of underlying models. Different difficulty models were needed for different question types, suggesting that people used a variety of approaches and, at a wider level, that both mental models and mental rules may be used in reasoning.

The psychological study of deductive reasoning has a long and productive history. However, research on deduction has been limited to a small number of artificial tasks, primarily syllogistic reasoning, conditional reasoning and Wason's selection task (see Evans, Newstead & Byrne, 1993, for a review). In the present studies, we extend this research by investigating a type of deductive task called 'analytical reasoning' (AR), which until recently was used in the Graduate Record Examination (GRE). This examination has been widely used to help in the selection of candidates for Graduate Schools

Correspondence should be addressed to Stephen E. Newstead, Centre for Thinking and Language, Department of Psychology, University of Plymouth, Drake Circus, Plymouth, PL4 8AA, UK. Electronic mail may be sent to: snewstead@plymouth.ac.uk

This research was supported by a series of research grants from Educational Testing Service, Princeton, NJ. The authors would like to thank Isaac Bejar, Pete Brittingham, Bill Eilfort, Kevin Singley, and David Williamson for their support and advice on the research.

http://www.psypress.com/tar DOI: 10.1080/13546780542000023

in the USA and has been shown to have predictive validity for that purpose (Kuncel, Hezlett, & Ones, 2001). While AR problems have at the moment been dropped from the current GRE, discussions are taking place about resurrecting them and they continue to be used in the LSAT (Law School Admission Test).

AR items involve a series of rules relating to a spatial array and are traditionally presented within a realistic scenario. They are deductive reasoning problems, in that they can be solved entirely on the basis of the information presented and have solutions that can be verified by formal logic. However, they are considerably more complex than the tasks traditionally used by psychologists in deductive reasoning research and it can be argued that they are consequently more representative of real world reasoning. It is also a matter of theoretical and practical interest in its own right to attempt (a) to understand the mental processes involved in such widely used tests of cognitive ability and (b) to predict what makes some items more difficult than others.

We address these aims by applying theoretical understanding of deductive reasoning processes gained from research on the standard paradigms mentioned above. Thus the present research can be seen as a partial test of the generality and practical applicability of theories developed in this field. We do not claim that it constitutes a fully-fledged test of current theories; nevertheless, an important perspective on the validity of the theories is provided by examining how well they cope with much more complex problems than those they have previously been used to explain. The research also provides preliminary information about the types of strategies that people adopt.

The present research was completed as part of a large scale research project with two aims: to create a difficulty model for AR problems, and to create a computer program to generate novel AR problems. The present paper describes the main difficulty modelling experiments. The first stage of this research project was a detailed breakdown of AR problem types identifying all the problem variables. In order to understand the present research it is necessary to explain in a little detail the nature of these problems.

The structure of AR problems

AR problems are designed to measure general analytical and logical reasoning abilities, and involve drawing inferences from presented information. A fairly typical instance is presented in Table 1 (taken from the ‘Big Book’ published by Educational Testing Service, 1996). The actual problem is presented in the left hand column and on the right hand side is a description of the structure and properties of the problem.

TABLE 1
A typical analytical reasoning problem

AR problem	*Structural analysis of AR problem*
An office building has exactly six floors, numbered 1 through 6 from bottom to top. Each of exactly six companies F, G, I, J, K, and M must be assigned an entire floor for office space. The floors must be assigned according to the following conditions:	Initial scenario (720)
F must be on a lower floor than G.	Initial rule 1: aboGF (360, 360)
I must be either on the floor immediately above M's floor or on the floor immediately below M's floor.	Initial rule 2: adjIM (240, 120)
J can be neither on the floor immediately above M's floor nor on the floor immediately below M's floor.	Initial rule 3: nadjJM (480, 96)
K must be on floor 4.	Initial rule 4: latK4 (120, 16)
1. Which of the following is an acceptable assignment of companies to floors, in order from floor 1 through floor 6 ?	Item 1: Possible order item. 16 models
(A) F, I, G, K, J, M	F (2,3)
(B) G, I, M, K, F, J	F (1)
(C) J, F, G, K, I, M	T
(D) J, M, I, K, F,	F (3)
(E) K, F, J, G, M, I	F (4)
2. If G is on floor 5, which of the following must be true ?	Item 2: Necessity item. Stem rule: latG5. 4 models
(A) F is on floor 1.	latF1 (F, 2 models)
(B) F is on floor 3.	latF3 (F, 2 models)
(C) I is on floor 1.	latI1 (F, 1 model)
(D) J is on floor 6.	latJ6 (T, Rules 1, 2, 4, stem)
(E) M is on floor 2.	latM2 (F, 2 models)
3. If M is on floor 2, any of the following could be true EXCEPT:	Item 3: Impossibility item. Stem rule: latM2. 4 models
(A) F is on floor 3.	latF3 (T, 2 models)
(B) F is on floor 5.	latI1 (T, 2 models)
(C) I is on floor 1.	latI1 (T, 2 models)
(D) J is on floor 5	latJ5 (T, 2 models)
(E) J is on floor 6.	latJ6 (T, 2 models)
4. If F and I are on floors one of which is immediately above the other, which of the following could be on floors one of which is immediately above the other ?	Item 4. Possibility item. Stem rule: adjFI. 4 models.
(A) F and J	adjFJ (F, Rules 2, 4, stem)
(B) F and M	adjFM (F, Rules 2, stem)
(C) G and M	adjGM (F, rules 1,2, 4 or Rules 2, 4, stem)
(D) I and K	adjIK (F, Rules 2, stem)
(E) J and K	adjJK (T, 2 models)

The first part of the problem in Table 1 is the initial scenario. This indicates the elements to be used in the problem (in Table 1, the companies) and the slots into which they have to be placed (in Table 1, the different floors). This opening scenario describes a vertical array with 720 possible orderings of the items (6*5*4*3*2) and this is the number given in the right hand column.

Following the scenario, an initial rule set is presented that places restrictions on which elements (companies) can be placed into which slots (floors). In Table 1 there are four initial rules and the right hand column gives a description of these. The first is an 'above' rule (abbreviated 'abo') indicating that one element is above another (the rule itself actually uses the word 'lower' but this is ignored in the abstract description of the rule). This rule on its own eliminates half of the possible orderings and so there are 360 remaining orderings (the figure given in the right hand column next to this rule). The second rule we term an 'adjacent' rule (abbreviated 'adj'). On its own this eliminates 480 orderings, leaving 240, and in combination with rule 1 leaves just 120 orderings; these are the figures given in the right hand column. Rule 3 is a 'negative adjacent' rule ('nadj') which leaves 480 orderings when applied on its own and 96 when applied in conjunction with the previous two rules. The fourth rule is an 'assign' rule (abbreviated 'lat' for 'letter at'), which leaves 120 orderings on its own and 16 when applied with the other three rules.

There then follow the actual test items which can come in different forms, though they always present five options of which there is always just one correct response. Item 1 in Table 1 is a possible orders item, in that respondents are asked to indicate which order could be true in light of the rules presented. The correct answer is option C. The other four choices all breach at least one of the rules. Option A breaches both rule 2 and rule 3, option B breaches rule 1, and so on (the rules breached are given against each option in the right hand column of Table 1).

Item 2 is what we term a necessity item in that it asks which of five options must be true in the light of the rules given. Note that this item is preceded by a further rule which applies just to this test item; such additional rules are called stem rules. This rule reduces the number of possible orderings to just four. Option D is the correct choice, since it is true in all four remaining orderings. The other options are false in at least one of the orderings, and the structural analysis indicates in how many models they are true (options A, B, and E are true in two of the four models, option D in just one). The structural analysis for the correct option indicates the minimum number of rules that are required to prove that it is the correct choice, in this case rules 1, 2, 4, and the stem rule.

Item 3 is an impossibility item, where the correct answer is the one statement which cannot be true. Option B is the correct choice since it is false in all of the four orderings remaining after the stem rule. All of the other options are true in two of the orderings. The rules given against the correct option are the fewest rules required to prove the option impossible.

Item 4 is a possibility item where the correct response is the only statement that could be true. The correct option, E, is true in two of the four orderings that remain after the stem rule. The other options are all definitely false, and the structural analysis indicates the minimum number of rules needed to prove each of them false. Note that the rules needed to prove the answer true are not listed as it always requires all the rules to logically determine possibility.

In the original problem there were seven test items, but since three of these used types of test item which have already been described they are not included in Table 1. The AR set presented in Table 1 is an order set, in that it involves placing the elements into slots which have a specific order and where the rules place restrictions on which sorts of orders are possible. The example was a vertical order; however, the same problem could be written with a scenario describing a horizontal or a temporal order. There are other types of AR problems but these were not studied in the present research.

Possible sources of difficulty

The central issue addressed by the present research concerns the sources of difficulty in such problems. Reasoning research suggests a number of possible contenders. We have divided these potential difficulty factors into four categories: firstly, possible influences of the content used; secondly, factors relating to the linguistic complexity of the rules; thirdly, factors relating to the difficulty of representing and manipulating models of the problem arrays; and fourthly, factors relating to different item types.

Content effects. It is possible that the content and context used to frame the problems might impact on difficulty. There is a wealth of evidence that people reason in different ways with different types of realistic material (see Evans, Newstead & Byrne, 1993, for multifarious examples of this). However, the scenarios used in AR sets are typically fairly abstract, often using letters and arbitrary categories such as boxes or positions at a table. Hence it seems quite likely that these will not lead to major differences in difficulty. It is also worth noting that effects of realistic material have not been reported with transitive inferences (e.g., three-term series problems) with anything like the frequency observed in other problems (though see Roberts & Sykes, 2003, for an exception to this).

The arrays described in the scenario can be vertical (above/below), horizontal (left to right or front to back), temporal (before/after), or based on any other form of ordering (e.g., taller/shorter). There is little evidence in the literature that this has an impact on difficulty (see Evans et al, 1993; Vandierendonck & De Vooght, 1996), and we initially treated these as equivalent (all were coded as the 'abo' rule, see Table 1). It is, however, an empirical matter (addressed in Experiment 3a) as to whether these differences are important in contributing to overall difficulty.

Complexity of rules. The number and complexity of the rules in a problem might be expected to affect difficulty, as might the difficulty of combining these rules. One of the most widely researched areas is that of negation, and there is overwhelming evidence that negation adds difficulty to simple logical rules (see Evans et al, 1993, for numerous examples of this). We also know that compound rules such as disjunctives can cause difficulty in processing. The exact level of difficulty is likely to be influenced by whether the disjunction is exclusive (e.g., 'Either A is in position 1 or B is in position 4, but not both'), or inclusive (e.g., 'Either A is in position 1 or B is in position 4, or both'). Realistic exclusive disjunctions have been found to be easier than inclusive ones (e.g., Newstead, Griggs, & Chrostowski, 1984).

Other compound rules might be expected to cause even greater difficulty. Conditionals are notoriously difficult to process, and we know that they lead to a number of common errors (see Evans & Over, 2004). People often interpret conditionals as biconditionals: they might regard the statement 'If A is in position 1 then B is in position 3' as also implying that 'If B is in position 3 then A is in position 1.' Such interpretations will lead to errors on some AR problems. We also know that a number of erroneous inferences are often made with conditionals. One of these is the denial of the antecedent inference: People who are told that the antecedent is false (i.e., A is not in position 1) infer (incorrectly) that the consequent is false (B is not in position 3).

There is thus good reason to believe that some of the rules used in AR problems will be more difficult than others on both linguistic and conceptual grounds. This difficulty may be evident in reading times, though this effect seems unlikely to be a major determinant of overall problem difficulty. More important may be the way in which the rules need to be used. For example, rules need to be combined in certain ways to determine the validity of the correct response and to exclude incorrect distracters. It seems probable that the more rules that are needed to do this and the more complex those rules are, the more difficult the problem will be.

It is worth pointing out at this stage that we are using the term 'rule' to refer to the propositions used to describe the restrictions on the orderings. This should not be confused with rule theories (e.g. Rips, 1994). Rule theorists claim that we have an internalised set of rules, a mental logic, which enables us to carry out reasoning tasks. If rules in our sense are found to predict difficulty, this does not commit us to a rule based approach; such a finding could be consistent with both rule theory and other approaches.

Difficulties of representation. A number of possible predictors relate to the complexity of the underlying models needed to represent AR problems. A highly influential theory of reasoning is the mental models theory, which claims that people construct mental representations of the situations described and draw conclusions from these (Johnson-Laird & Byrne, 1991). In this theory, the number of different models that need to be constructed is a prime determinant of difficulty. Hence one might predict that the number of models which are possible following the initial rule set (and the stem rule, if there is one) will be related to the difficulty of AR problems. Clearly, it would be unrealistic to expect that people analyse all of the models in any AR set since these usually number in the hundreds, but the number of models typically becomes more manageable after all the rules have been presented, and so this may be a potential predictor. It is also possible to construct partial arrays that can be used to represent possible model information in a more concise format.

In AR problems there are a number of potentially relevant model factors. The number of models possible in the initial scenario might be a factor, as might the number possible after the initial rule set has been represented, or the number possible after the stem rule.

Test item factors. Another potential predictor is the type of test item used, and reasoning research has produced evidence relevant to the difficulty of necessity, possibility and impossibility items. It might be expected that it would be easiest to determine whether a conclusion is possible, since it is only necessary to find one situation in which that conclusion holds in order to be certain that it is possible. However, all of the situations need to be searched in order to be sure that a conclusion is necessary (i.e., true in all models). Similarly, all models need to be searched in order to determine that a conclusion is impossible (i.e., false in all models). Such findings have been obtained both in studies of propositional reasoning (Bell & Johnson-Laird, 1998) and in studies of syllogistic reasoning (Evans, Handley, Johnson-Laird, & Harper, 1999).

However, this conclusion needs to be tempered since accepting the right conclusion (the key) involves rejecting the incorrect distracters (the four false options) and the nature of these is likely to be important. For example, with necessity items the distracters might be possible (i.e., true in some but not all situations), or impossible. If the distracters are all impossible (i.e., they do not hold in any situations) then identifying the key should be relatively straightforward since it is the only option for which there is a true model. In contrast, if one or more of the distracters is possible it will be more difficult to identify the key since a more thorough checking of the remaining model set will be needed. The structural analysis in Table 1 indicates the number of models in which possible test items hold, both for correct (key) and incorrect (distracter) items.

Pilot study

To supplement the insights provided by previous research we carried out a small pilot study of people thinking aloud while solving AR problems. The complete version of the companies and floors problem (see Table 1) was presented to eight participants who were allowed as long as they needed to solve the items while continuously verbalising what they were thinking. There were a number of common themes in the way that the participants approached the problem. All of them first familiarised themselves with the scenario, and reported using some kind of graphic representation of the rule implications using a vertical array. They all tackled the last rule (K must be on floor 4) first, by indicating on their array that position 4 contained company K. They then tackled the first rule, that F is below G, by using some kind of graphic representation, usually just placing the F below the G to one side of the vertical array. The two remaining rules were always tackled last, and several participants combined these, which was possible because both rules refer to company M. The negatives in these rules were represented in different ways, sometimes by a large 'X' next to them, sometimes by the word 'no' next to them.

This preliminary study of think-aloud protocols provided strong clues as to the relative difficulty of some of the rules. It was clear that the 'assign' rule had some priority, being dealt with first even though it was the fourth and final rule in the initial rule set. The 'adjacent' rule was dealt with next, and again participants found it relatively easy to handle. The two disjunctive rules seemed to be the most difficult. Negation seemed to involve a further processing step. These findings in general lend weight to the conclusions suggested by previous research.

Some of the findings add to our understanding. It was a little surprising to us that participants did not consider the rules in the order presented but invariably started with the 'assign' rule. This may reflect a tendency to start with the most informative premise (i.e., the one which eliminates the most possibilities), as claimed by Johnson-Laird, Byrne, & Schaeken (1994). However, there is clearly more to it than this since the rule processed next, rule 1, is actually less informative than rule 2. There are also indications as to the representation used, since participants tended to construct something akin to a spatial mental model. However, this is clearly only a partial model, with some fixed elements and a number whose position is left ambiguous. For this reason we investigated model complexity as a predictor of difficulty in addition to the overall number of models. The measure we developed, the model variability score, is essentially an indication of the range of positions elements may occupy in the model set.

Two main strategies seemed to be used by participants, which can be characterised as rule checking and option elimination. The rule checking strategy was used with possible orders items. It involved either systematically checking each option against each of the rules or systematically checking each rule against each option. The option which was not ruled out by this process was the correct choice. For test items other than possible orders, the more complex option checking tended to be used. To illustrate, on necessity items each option would be checked to see if there was a model that could be constructed which invalidated it. If so, it was eliminated as a distracter. If no such model could be found, the option was flagged as a possible correct answer and the participant then continued with the other options. If more than one option remained at the end of this process, further checking was carried out. One might surmise that this elimination strategy would be more efficient where there are relatively few models to consider or where the models are similar to each other.

There were two main stages in the present research. The first stage (Experiments 1 and 2) involved experimental studies of the difficulty of various components of order sets, principally the rules involved. These are the building blocks of AR problems and we assumed that their difficulty would be central to overall problem difficulty. Based on these findings, we developed a difficulty model and tested this against new data collected in Experiments 3a and 3b.

EXPERIMENT 1

The principal purpose of Experiment 1 was to determine the relative difficulty of simple rules used in AR problems. The task used to assess difficulty was a simple true/false verification task. Participants were presented with an order scenario describing six elements that had to be placed in six slots. There followed a single rule and then an ordering of the six elements. Participants were required to indicate whether this ordering was true or false with respect to the initial rule presented. The rules used, which are among the most commonly used rules in AR order sets, are listed in Table 2. Note that the final four rules are negations of rules also used in this study (the rules ‘above’ and ‘below’ are never negated in AR

TABLE 2
Reading times, answer times and error rates in Experiment 1

Rule	*Semantic informativeness*	*Mean reading time (seconds)*	*Mean answer time (seconds)*	*Mean % error*
Assign	600	1.91	1.70	3.87
Immediately above	600	3.35	2.27	7.74
Immediately below	600	3.71	2.21	5.16
Adjacent	480	4.55	2.40	3.55
Above	360	3.12	2.19	4.84
Below	360	3.38	2.28	6.13
Not adjacent	240	5.28	2.86	7.74
Negative assign	120	2.47	2.12	7.74
Not immediately above	120	4.75	3.25	10.65
Not immediately below	120	4.80	3.07	9.35
Correlation with semantic informativeness		−.41	−.68	−.72

sets since ‘not above’ is the same as, and more easily expressed as, ‘below’).

Method

Participants. Participants were 31 undergraduate or postgraduate students of the University of Plymouth who were paid for their participation.

Materials and procedure. The experiment was computer based and utilised the Psyscope experiment generating software developed by Cohen, MacWhinney, Flatt, and Provost (1993). Responses were made using the computer keyboard which had been adapted to include ‘yes’ and ‘no’ keys. These were counterbalanced so that half the participants had the ‘yes’ key on the left of the keyboard and the ‘no’ key on the right, the other half having these key positions reversed.

Participants were instructed that the same scenario applied to all the experimental trials and related to the location of offices in a building (it was actually the scenario presented in Table 1 but using the letters A, B, C, D, E, and F). They were told a rule would appear in the bottom half of the screen, and that when they were ready they should press the space bar. This would bring up an order of offices, and they were asked to indicate by pressing the ‘yes’ or ‘no’ key whether or not this was an acceptable ordering given the rule they had been presented with.

These instructions were followed by four practice trials after which participants had an opportunity to ask questions before the experiment proper. During the experiment participants were given the opportunity of taking a break after each block of experimental trials.

In each trial participants were presented with the scenario followed by a rule such as ‘C is on a floor immediately above E’. Each rule was presented ten times using different content, making 100 experimental trials in total. These trials were presented in random order in blocks of 20 trials. When participants had read the rule and were ready to proceed they pressed the space bar, which enabled reading times to be measured for each rule. A possible order was then given, for example AEDFCB, and the time it took to respond was recorded; this was the answer time. On half the trials the rules were true of the ordering and on half they were false. ‘Assign’ rules never assigned a company to one of the end positions in the ordering (i.e., the top and bottom floors) since there is evidence that end items are easier to process (Potts, 1974) and they tend to be avoided by AR item writers. Apart from this restriction, the elements included in the rule and the order of the elements in the false cases were selected at random.

Results

Reading times are presented in Table 2. They were analysed using a one-way analysis of variance which revealed a significant effect of rule type, $F(9, 270) = 33.59$, $MSE = 2.22$, $p < .01$, in which ‘assign’ rules were read faster than any other rule form. As there were four rules with direct negative counterparts it was possible to analyse the data from these eight rules in a two way analysis of variance with positive vs. negative rule type as a factor. The results of this analysis showed that positive rule forms were read significantly faster that their negative counterparts, $F(1, 30) = 24.64$, $MSE = 4.47$, $p < .01$.

Answer times are also presented in Table 2. A rule (10) by truth status (true vs. false) analysis of variance revealed a main effect of rule type, $F(9, 270) = 17.34$, $MSE = 0.81$, $p < .01$, a main effect of truth status, $F(1, 30) = 21.12$, $MSE = 0.50$, $p < .01$, and a significant interaction between these two factors, $F(9, 270) = 2.09$, $MSE = 0.53$, $p < .05$. The source would seem to correspond to a well established finding in the literature. As can be seen in Figure 1, on affirmative rules (the first six in Figure 1) true responses are easier than false ones, but this is not the case for negative rules (cf., Wason, 1959; Clark & Chase, 1972).

Overall error rates were low (mean 6%) and are presented in Table 2. There was a main effect of rule type, $F\ (9, 270) = 2.81$, $MSE = 0.01$, $p < .01$. Most of the errors were made by just three participants, which precluded any meaningful further analysis.

The semantic informativeness of each rule (i.e., the number of possible orders that rule eliminates) was calculated and is presented in Table 2. As

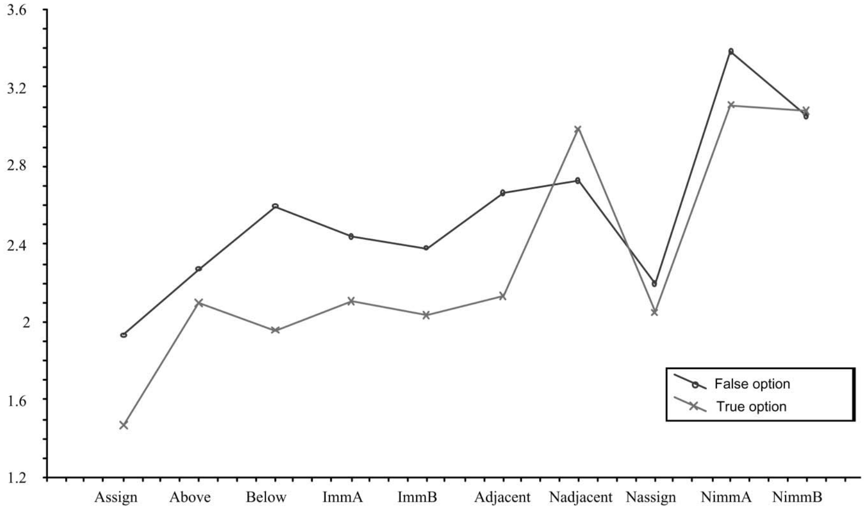

Figure 1. Interaction on response times between rule type and truth in Experiment 1a.

can be seen in the bottom row of this table, there were reasonably high negative correlations between semantic informativeness and difficulty on the various measures, all of which were statistically significant ($p < .05$). However, these overall correlations mask some marked divergences from this, for example that 'assign' rules are so much easier than the equally informative 'immediately above' and 'immediately below' rules.

Discussion

This experiment has provided a detailed picture of the overall difficulty of rules. As a rough approximation, it would appear that there are three groups of rules in terms of overall difficulty. 'Assign' rules are in a group of their own, being easier than any other rules, followed by a group containing 'immediately above', 'immediately below', 'adjacent', 'above', 'below' and 'negative assign' rules, and finally a group containing 'negative adjacent', 'negative immediately above' and 'negative immediately below' rules. It is clear that, as expected, negation contributes to difficulty: in every case in this experiment, negated rules were more difficult than their affirmative counterparts on all measures.

In order to check on the consistency of these findings, we carried out two follow up verification studies using three of the rules: 'assign', 'adjacent' and 'above'. In one study we used these in combinations of two rules and in another we used all three rules. In both studies, 'assign' rules were significantly easier than the other two, which did not differ from each other.

Although this is not a complete replication it indicates that our findings are robust.

This study examined only simple rules. Compound rules are those that combine two simple rules, for example by combining two ‘assign’ rules into a conditional rule such as ‘If A is in position 1 then D is in position 4’. The next study investigated the difficulty of such rules.

EXPERIMENT 2

This study used four compound rules: conditionals, biconditionals (e.g., ‘If D is in position 1 then C is in position 3 and if C is in position 3 D is in position 1’), inclusive disjunctives (e.g., Either C is in position 2 or D is in position 4, or both), and exclusive disjunctives (e.g., Either B is in position 1 or C is in position 2, but not both). Compound rules combine two simple rules and we varied the simple rules contained in compound rules so that some used ‘assign’, some ‘immediately above’, some ‘adjacent’, and some ‘above’ rules. These compound rules were presented as part of an initial rule set containing three other rules. We already know that compound rules are more difficult than the simple rules used in Experiment 1 but we do not know how much more difficult. Furthermore, little is known about the difficulty of compound rules relative to each other, and indeed the relative difficulty of disjunctive rules seems to vary with the material used (see Evans, Newstead & Byrne, 1993). Hence it is important to assess difficulty in situations analogous to AR problems.

Method

Participants. Ninety-nine undergraduate students from the University of Plymouth were paid for their participation in the experiment.

Materials and procedure. Participants were given a booklet containing a standardised instruction sheet and sixteen questions with each question on a separate sheet. Each of the sixteen questions included a compound rule. There were four questions with one conditional rule, four with one biconditional rule, four with one inclusive disjunctive rule, and four with one exclusive disjunctive rule. For each of the four different types of compound rules, there was one with an ‘assign’ sub-rule, one with an ‘above’ sub-rule, one with an ‘immediately above’ sub-rule and one with an ‘adjacent’ sub-rule. (Sub-rule is the type of simple rule upon which the compound rule is based). The order of the trials was fully randomised across participants. Following the rules, participants were presented with a possible order item in which they had to indicate which of five presented orders conformed to the rules. Each of the alternatives (the distracters) broke just one of the rules in the initial rule set.

The instructions described the form of the problems, ensured confidentiality, and emphasised that only one of the five options was correct. Instructions also told participants they would have a limited time to answer each question and that 20 seconds before the end of the time the experimenter would warn them of the ending of the allocated time. Participants were informed they could make notes during problem solution, as is possible in the delivery of actual AR sets. Once they had asked any questions arising from the instructions they were asked to start the questions. Participants were given 90 seconds to answer each of the questions, which corresponds to the time notionally allowed in the standard administration of the GRE.

An example may illustrate the procedure more clearly. This was a problem involving a conditional rule:

> Each of exactly six objects – A, B, C, D, E and F – is to be placed in one of six slots arranged horizontally and numbered 1 to 6 from left to right. Each slot must have one of the objects placed in it. The arrangement of these objects in the six slots is subject to the following conditions.
> A must be placed in a slot somewhere to the left of B
> C must be placed in slot 4
> A must not be placed in slot 3.
> If E is placed in slot 1 then D must be placed in slot 3.
> Which of the following arrangements of objects, from slot 1 to slot 6, conforms to the rules?
> D E A C F B
> E A F C D B
> F E B C D A
> A B F D C E
> A E B C D F
> The conditional rule is the final one in the initial rule set and combines two ‘assign’ sub-rules. The correct option is (E). Each of the other alternatives breaks exactly one of the rules in the initial rule set: (A) breaks the third rule, (B) breaks the fourth rule, (C) breaks the first rule, and (D) breaks the second rule.

Results and discussion

We had two measures of the difficulty of compound rules. The first of these was the overall difficulty of the set containing each compound rule. This is a fairly crude measure since overall difficulty will be affected by the other rules in the initial rule set and these varied from one initial rule set to another. Nevertheless, we can still draw some tentative conclusions. The analysis of errors found a significant effect of compound rule type, $F(3, 294) = 10.72$, $MSE = 0.13$, $p < .01$, and a breakdown analysis revealed that biconditional rules attracted twice as many errors as the other compound rule types. There was also a significant effect of sub-rule type, $F(3, 294) = 16.82$, $MSE = 0.11$, $p < .01$. Compound rules with 'assign' sub-rules were the easiest, followed by those with 'immediately above', 'adjacent' and then 'above' sub-rules. There was also a significant interaction between these two factors, $F(9, 882) = 4.68$, $MSE = 0.13$, $p < .01$. The nature of these effects is shown in Figure 2.

The other, and arguably more accurate, measure of the difficulty of compound rules was the number of errors made on each distracter type. As indicated above, each distracter breached just one of the rules in the initial rule set, and so the frequency with which that distracter was chosen as the (incorrect) alternative gives a measure of how difficult the associated rule was. The error rates associated with the distracters breaking each of the different rule types are presented in Table 3. An analysis of variance revealed a significant effect of rule type, $F(7, 48) = 5.52$, $MSE = 19.35$, $p < .01$. Inclusive disjunctives were the easiest, followed by exclusive disjunctives and conditionals (which were not significantly different from each other), and with biconditionals the most difficult of all.

Based on the studies of both simple and compound rules, we were in a position to derive a simple difficulty metric. The weightings are presented in Table 4. 'Assign' rules were defined as having a weighting of one, and the other rules were allocated weights as multiples of this. These

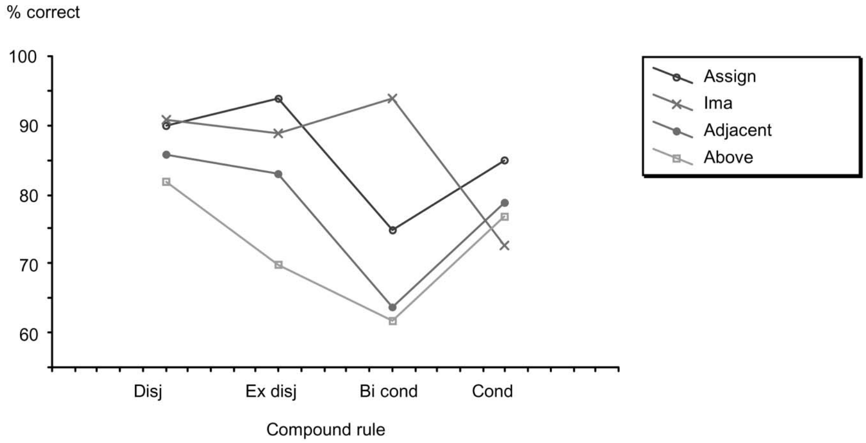

Figure 2. Interaction between compound rule type and sub-type in Experiment 2.

TABLE 3
Mean errors associated with distracter types in Experiment 2

Rule broken	*Mean % error*
Assign	2.1
Immediately above	3.1
Adjacent	3.7
Above	3.9
Inclusive disjunctive	4.8
Exclusive disjunctive	8.1
Conditional	8.1
Biconditional	16.4

TABLE 4
Preliminary rule difficulty weightings

Rule type	*Weight*
Assign	1
Immediately above	2
Adjacent	2
Above	2
Negative assign	2
Inclusive disjunctive with assign rule	2
Conditional with assign rule	3
Inclusive disjunctive with above, adjacent or immediately above rule	3
Exclusive disjunctive with assign rule	3
Exclusive disjunctive with above, adjacent or immediately above rule	4
Conditional with above, adjacent or immediately above rule	4
Negative with above, adjacent or immediately above rule	3 (4)
Biconditional with assign rule	4
Biconditional with above, adjacent or immediately above rule	8

Conjunctive rules effectively consist of two separate simple rules, and difficulty weightings reflect the sum of the weights of the two rules in question. The figure in parentheses is the revised weight after testing in Experiment 3a.

weightings are based primarily on the results of Experiments 1 and 2, but we also drew on previous research and our own intuitions. They are rounded to the nearest whole number and hence are approximations rather than exact estimates.

DEVELOPMENT OF THE DIFFICULTY MODEL

The next stage of the research involved generating a difficulty model so that its effectiveness could be tested against new data. We have not studied all of the factors that might potentially contribute to difficulty but there is enough information to begin developing at least a preliminary difficulty model for these AR sets.

We now have approximate weightings for the rules used, as presented above in Table 4. A simple possible predictor is the combined weightings of the rules used in the initial rule set and stem rule (if any). For example, an initial rule set with an ‘assign’ rule (weight 1), an ‘adjacent’ rule (weight 2) and a ‘negative adjacent’ rule (weight 4) will have an overall initial rule set difficulty of 7. In addition, if there is a stem rule then this will add further difficulty. It is possible that the weights of the initial rule set and the stem rule can simply be added together to provide a predictor but it is also possible that they act as independent predictors. The difficulty of the rule used in the options might also influence difficulty.

Another potential predictor is the number and complexity of the rules that need to be used in order to determine that the key is correct and to eliminate the distracters. To illustrate, Table 1 presents a listing of the rules needed with the offices and floors problem. A rule needs score can be derived by summing the weights of the rules in the rule combinations minimally needed to prove the key and to eliminate the distracters. Where two different combinations of rules can be used to prove an item true or false, the one producing the lowest weighted score was used in the rule needs score.

There are a number of other potential predictors relating to the number and complexity of the group of valid models that remain after all the rules have been applied. A simple measure is the number of models which remain after application of the initial rule set or after application of the initial rule set and the stem rule. We also derived a measure of the variability of those models. Consider, for example, these two pairs of models:

(i) A B C D E F
(ii) B A C D E F

(iii) A B C D E F
(iv) C E A F D B

The first pair may be easy to encode since the only difference is in the order of the first two elements. In the second, however, there is much more variability and this pair will presumably be less easy to encode and reason with. Hence, in addition to the overall number of possible models we also calculated a model variability score. This is simply the sum of the number of different elements that can be in each position. In the example given above, for the first pair two elements can be in position 1, two in position 2 and only one in each of the other four positions. Hence the total model variability score is 8. In the second pair, there are two possibilities in each of the six positions and hence the model variability score is 12. This measure is in essence an indication of how many fixed elements there are and how much variation is possible with the non-fixed elements.

Another model-related measure concerns the number of models in which answer options are true. By definition a necessary option is true in all the models remaining after the stem rule and an impossible option is true in none of them. However, possible options can be true in different numbers of these models. Possible answer options necessarily occur as the key in possibility items and as the distracters in impossibility items. They may also occur as the distracters in necessity items. It may be easier to identify distracters to necessity items if the distracters are impossible rather than possible as there are no models in which an impossible distracter is true.

In addition, the proportion of models in which a possible option is true may be important. For example, if there are ten remaining possible orderings following the initial and stem rules, a possibility key might be true in only one of these models or it might be true in nine of them. In the first instance the search for a true case would be more difficult (a 1/10 chance of success with each model generated) than in the second instance (where there is a 9/10 chance of success with each model generated). This factor works in the other direction where a possible option is a distracter in a necessity item. In this case the option can only be correctly identified as a distracter when a false case is found (this proves that the option is not the necessary key). Thus, a higher proportion of false cases will make this identification easier. Hence, another potential predictor of difficulty is the proportion of models in which possible keys and distracters are true. This potential difficulty factor might be expected to influence different item types in different ways, which would result in different difficulty models for each of the different item types.

The purpose of Experiment 3 was to test these and other predictors more formally and fully. The test items were generated by the program we had developed (see Bradon, Evans, Handley, Newstead, Dennis & Prat-Sala, undated), and this also produced values for all the potential predictors (see structural analysis in Table 1).

EXPERIMENT 3

Experiment 3 was designed to create and validate models of difficulty for each of the item types in simple order sets. Specifically, Experiment 3a was designed to validate the weightings for different kinds of simple rules generated in previous experiments, to determine the appropriate weightings for factors shown to influence difficulty in previous experiments, and to determine the possible influence of factors not previously considered. The experiment was designed to produce a regression analysis of difficulty factors for each of the items from which potential difficulty models could be constructed. Experiment 3b was designed to validate the model produced by Experiment 3a. This experiment was run at the same time as Experiment 3a and presented a different but similar group of AR sets to different participants from the same population.

EXPERIMENT 3a

Method

Participants. Participants were 122 sixth form students from Ivybridge Community College, Devon, UK, aged 17 or 18 years. In the British system compulsory education finishes at age 16, but the more able students can continue for a further two years in the sixth form (Years 12 and 13) where they will take further exams as the basis for university entrance or other qualifications. A high proportion of sixth form students are expected to continue into university education. Although AR items were developed for use with university students, the results vindicated our selection of this group of participants since their performance was, if anything, better than that of university students we tested on similar problems. Participants did not receive individual payment but the College sixth form was paid a total sum for the participation of all the students in both parts of Experiment 3.

Materials and procedure. The experiment was a pencil and paper based task in which each participant was presented with detailed instructions followed by 15 AR sets, each with four test items. Each AR set had a possible orders item, a necessity item, a possibility item and an impossibility item. The instructions contained an introduction to the experiment, an example of an AR item, a set of strategy suggestions (taken from the Big Book introduction to AR items) and instructions on timing and responses.

The 15 four-item sets used in the experiment were chosen from a group of 60 sets generated by the AR generation program we developed. All sets used six elements. Each set was clothed in three different scenarios representing vertical (above/below), horizontal (left to right or front to back), or temporal (before/after), orderings. There is little evidence in the literature that the type of ordering used has an impact on difficulty (see Evans et al., 1993; Vandierendonck & De Vooght, 1996), but the design of this study allowed this possibility to be investigated. Scenario type was counterbalanced between participants and item order was randomised. Sets were specifically chosen to vary in terms of a wide range of different potential difficulty factors whilst avoiding high correlations between the factors. The

TABLE 5
Metrics relating to potential item difficulty

Comprehension factors	Number of initial rules
	Initial rule score Weighted total score for the initial rule set
	Stem rule score Weighted score for the stem rule
	Item rule score Total weighted score for the stem rule the initial rule set
	The weighted score of the option type used in the item
Model-based factors	Initial models Number of possible orders after application of the initial rule set
	Item models Number of possible orders relating to the item
	Number of models in which a given option was possible
	Proportion of models in which a given option was possible
	Model variability score a variability metric for the item model set (NB For Possible order items this is the variability of the initial model set)
Rule-based factors	Key rule needs score The weighted score of the minimum number of rules needed to confirm or falsify the key.
	Total distractor rule needs score The total of the weighted scores of the minimum number of rules needed to confirm or falsify each distractor.

factors taken into consideration can be seen in Table 5. The possible orders item was always presented first, with the order of the other three item types being counterbalanced across trials. It is normal practice in the GRE to present a possible orders item first since these are known to be relatively easy and thus serve as a useful introductory item, though this does mean that any comparison between these and other types of test item is confounded. An example of a clothed scenario is presented in Table 6.

Participants were allowed 7 minutes to answer the four test items associated with each scenario; this was slightly more generous than normal ETS timing to minimise errors caused by time pressure alone. They were warned when there was one minute left and when there were 20 seconds left, and were asked to make sure that they gave an answer to every item.

Results

Differences between conditions. Experiment 3a was designed primarily to generate a difficulty model based on the regression of variables related to set and item structure. In order to do this it was necessary to counterbalance for variables within set presentations. Each set was presented in three different clothed forms: as a vertical, horizontal and a temporal order set. In

TABLE 6
Example of an AR set used in Experiment 3

A woman plans to plant exactly six kinds of herbs: marjoram, oregano, parsley, rosemary, sage, and thyme. She places six pots side by side in a straight line and numbers the pots 1 to 6 consecutively from left to right. She will plant only one kind of herb in each pot. The arrangement of the herbs is subject to the following conditions:

Parsley must be planted in pot 4.
Rosemary must be planted somewhere to the left of Parsley.
Thyme must be planted immediately to the right of Oregano.

Which of the following arrangements of herbs from pot 1 to pot 6, conforms to the conditions above?

(A) rosemary, oregano, thyme, parsley, marjoram, sage
(B) sage, oregano, thyme, parsley, rosemary, marjoram
(C) marjoram, rosemary, oregano, parsley, thyme, sage
(D) rosemary, parsley, sage, marjoram, oregano, thyme
(E) thyme, marjoram, rosemary, parsley, sage, oregano

If Oregano is planted somewhere to the right of Sage, which of the following COULD be true?

(A) Parsley and Thyme are planted in adjacent pots.
(B) Oregano and Sage are planted in adjacent pots.
(C) Marjoram and Thyme are planted in adjacent pots.
(D) Thyme and Rosemary are planted in adjacent pots.
(E) Rosemary and Sage are planted in adjacent pots.

If Marjoram is planted somewhere to the right of Rosemary, ANY of the following COULD be true *EXCEPT*?

(A) Marjoram is planted in pot 2.
(B) Oregano is planted in pot 5.
(C) Thyme is planted in pot 5.
(D) Rosemary is planted in pot 1.
(E) Marjoram is planted in pot 6.

If Rosemary is planted somewhere to the right of Sage, which of the following MUST be true?

(A) Marjoram is planted in pot 3.
(B) Marjoram is planted in pot 1.
(C) Sage is planted in pot 2.
(D) Oregano is planted in pot 5.
(E) Rosemary is planted in pot 2.

addition, within each set, necessity, possibility, and impossibility items were presented in every possible order (following the possible orders item which was always presented first). As these factors were completely counterbalanced it was possible to perform analysis of variance to look for possible differences. There were no differences between the different scenarios $F(2, 363) = 0.014$, $MSE = 0.04$, $p = .99$. The means were remarkably close to each other, all rounding to 59% correct. This is an important finding in terms of developing difficulty models since it confirms the prediction that the type of transitive relationship used (vertical vs. horizontal vs. temporal) has little effect on difficulty.

Position within the set had an effect, $F(2, 242) = 6.53$, $MSE = 0.01$, $p < .01$ (this analysis excludes possible orders items which were always presented first). The mean number of correct responses was 51.8% in position 2, 48.3% in position 3 and 45.5% in position 4. Clearly, items presented later were answered less accurately. As the sets were time constrained (7 minutes for all 4 items) this result presumably reflects participants running out of time, but it does highlight the need to allow for the position of an item within the set when calculating overall item difficulty.

There was also a significant difference between item types, $F(3, 56) = 52.84$, $MSE = 109.86$, $p < .01$, which is presented in Figure 3. Clearly, possible orders items are by far the easiest, though recall that this effect is confounded by the fact that these items were always presented first. There is relatively little overall difference between possibility, impossibility and necessity items. In Figure 3 the difficulty of different types of necessity item is presented. In some of these, all of the distracters were impossible, in others 1, 2, or 3 distracters were possible. Follow up analysis demonstrated that necessity items with all impossible distracters were significantly easier than those with mixed distracters. This finding suggests a possible problem as the experiment was not designed to produce separate difficulty models for different forms of necessity item. Within each item the position of the key was varied. There were no significant differences between the five possible key positions, $F(4, 55) = 0.12$, $MSE = 424.80$, $p = .97$.

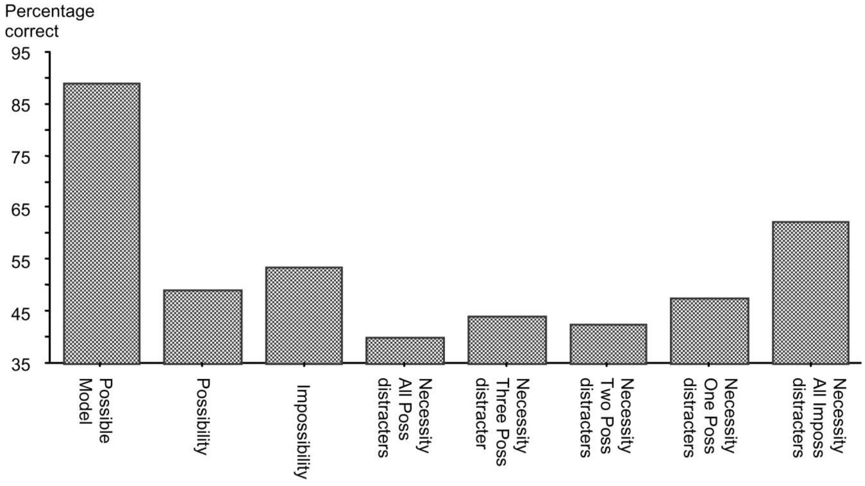

Figure 3. Mean scores for each item type in Experiment 3a.

Testing the difficulty model. The weightings in the initial rule difficulty index (see Table 4) were originally derived from the answer time scores in Experiments 1 and 2. These proposed weightings were tested by using them to model the expected pattern of errors on possible orders items. Comparison of the expected and actual error scores confirmed the majority of the weightings but led to one small change, an increase in the weightings for the 'not adjacent' and 'not immediately above/below' rules from 3 to 4. These revised weightings were the ones tested in the following analyses.

Separate regression analyses were performed on the percentage correct scores for the four different item types using the factors shown in Table 5. The experiment was designed to allow for this simple form of analysis by ensuring that all participants completed every item. Thus, each item score reflects the mean of the total participant group and is not biased by variation in respondent groups or respondent ability differences.

Possible orders items. Overall accuracy was high with a mean score of 89% correct. As many of the metrics listed above were not relevant to possible orders items, potential difficulty factors included in the regression were: number of initial rules; initial rule score; initial models; total distracter rule needs score; and model variability score (in this case model variability score reflects the variability of the initial model set).

The model resulting from a stepwise regression is presented in Table 7. Only one of the variables put into the regression, initial rule score, enters into this equation. The prime determinant of difficulty for these items thus appears to be the overall difficulty of the rules presented. Note that for possible orders items the initial rule score is equal to the key rule needs score in other items, since a possible key will always require the combination of all the rules to determine that it is not impossible. Although in this case the intercept is above 100, the equation will never return predicted scores of above 100% as sets must have some form of initial rules. The possible orders

TABLE 7
Predictive models for different item types

Type of item	Predictive model
Possible Orders	104.49 – (2.42 * Initial Rule Score)
Possibility	102.85 – (4.39 * Item Rule Score) – (.68 * Distracter Rule Needs Score)
Impossibility	110.15 – (Item Key Rule Scores) (5.73 * Option Rule Score) (1.19 * Model Variability Score)
Necessity (overall)	79 – (4.45 * Possible Distractors) (Item Key Rule Score) (.72 * Model Variability Score)

items difficulty model had an adjusted R^2 of .51. This model makes good intuitive sense since a plausible strategy for solving possible orders items – and one suggested by our pilot study using think aloud protocols – is to check each option against each rule; hence the overall difficulty of the rules would be expected to be a key factor.

Possibility items. Possibility items had an overall mean score of 49% correct. The regression equation for possibility items is presented in Table 7. For these items there were two main predictors. The first is the item rule score, which is simply the combined weight of the initial rules and the stem rule. The second is the distracter rule needs score, which is the sum of all the rule weights of the combinations of rules needed to eliminate each of the four distracters. There is intuitive appeal to this model since a plausible way of solving these items is to process all the rules and then use some of them to eliminate each of the distracters. The possibility items difficulty model had an adjusted R^2 of .72.

Impossibility items. Impossibility items had an overall mean score of 54% correct. The regression equation can be seen in Table 7. The predictive model involves the item rule score (the combined weights of all the rules), the key rule needs score (the combined weights of the rules needed to determine that the key is impossible), the option type rule score (the weighted score of the rules used in the options) and the model variability score. It is not surprising that all rules contribute to the model since distracters on impossibility items require the use of all rules in order to eliminate them. When this is added to the key rule score one is left with a measure of all the rules needed to work out that the distracters are the wrong choice and the key is the right one. The presence of option rule type in the model reflects the fact that problems which ask questions about assignment are easier than others on impossibility items. Model variability score is a measure of the extent to which the remaining models (after the stem rule has been presented) vary. This reflects the fact that impossibility items are more difficult when the model sets are highly variable, presumably because they are more difficult to represent. The impossibility items difficulty model had an adjusted R^2 of .43.

Necessity items. The overall mean score for necessity items was 47% correct, although as can be seen in Figure 3 there were large differences within necessity items depending on the nature of their distracters. The regression equation for these items can be seen in Table 7 and involves the item rule score and the key rule needs score, just as impossibility items did (and for similar reasons). Again as with impossibility items, the model variability score is a predictor. The other major predictor is possible distracters, which is a measure of the number of distracters that are possible (as opposed to impossible). It may well be that necessity items with possible, impossible and mixed distracter types should be treated as different cases, but unfortunately this was beyond the scope of the present experiment which only presented three examples of each necessity item type. The overall necessity items difficulty model had an adjusted R^2 of .64.

EXPERIMENT 3b

This experiment was designed to test the robustness of the models generated by Experiment 3a. For convenience it was carried out using the same participant population and was conducted simultaneously with Experiment 3a, but as indicated earlier there is good reason to believe that these participants are not unlike undergraduate students.

Method

Participants. Participants were 72 sixth form students from Ivybridge Community College, the same participant group as Experiment 3a. Participants did not receive individual payment.

Materials and procedure. The experimental materials and procedure were exactly the same as those for Experiment 3a except that 15 different AR sets were used. The experiment was run concurrently with Experiment 3a. Participants were randomly selected for Experiment 3a or 3b from the same population.

Results and discussion

By applying the item difficulty models generated by Experiment 3a, predicted scores were calculated for each item in the 15 sets presented in Experiment 3b. These scores were correlated with the actual mean scores recorded in Experiment 3b. The correlations are shown in Table 8.

TABLE 8
Correlations between predicted and actual scores in Experiment 3b

Item type	*Correlation*
Possible order	0.60
Possibility	0.54
Impossibility	0.80
Necessity	0.57
Overall	0.83

Regression using predicted scores as a variable gave an adjusted R^2 of .70. The models would thus appear to be robust.

GENERAL DISCUSSION

The main purpose of the present research was to develop and test a difficulty model for AR items used in the GRE. The model was developed in part using predictions derived from reasoning theory and research, but the main foundation for the model was our own research on the relative difficulty of the rules used in AR items. The importance of rule difficulty was demonstrated by the effectiveness of the models in predicting actual behaviour using metrics derived from rule weightings.

However, it is clear that a single model is not possible since the parameters for predicting different types of test item are different. Instead we have developed different difficulty models for each item type. In general these seem intuitively plausible, since they depend on such factors as the difficulty of the rules used in the initial rule set and the stem, the difficulty of the rules that need to be applied in accepting the key and rejecting the distracters, and the complexity of the models which remain after all rules have been applied.

Furthermore, these factors map on to and shed additional light on the strategies revealed in the pilot study. For possible orders items the difficulty of the rules was the main predictor, supporting the existence of a strategy of rule checking. The predominant strategy with these items seems to involve systematically going through the options and checking them against the rules, or alternatively systematically going through the rules and checking these against each option. For both necessity and impossibility items, model variability score was a predictor, supporting the claim that people try to construct models to support a key or reject a distracter. The more complex the models that need to be constructed, the more difficult this strategy will be to execute. However, things are more complicated than this since with both of these types of test item, and with possibility items, rule needs score entered into the regression model. This indicates that as well as constructing model representations, people also try to combine rules to draw logical conclusions from them.

What is not clear from our data is whether the difficulty model stems from different people using different strategies or from each person using a combination of strategies. In an attempt to differentiate between these two possibilities, we carried out a cluster analysis on the problems used in Experiment 3a. Two distinct clusters emerged, suggesting that some problems were easier for some people while different problems were easier for other people. However, there seemed to be no systematic differences between the items in each group. It was not the case, for example, that problems that were high in model variability fell into one group whereas those that contained high rules needs scores fell into another. Of the measures that went into our analysis, there was not one that mapped clearly onto either of the clusters. We did not have sufficient data to do any more detailed analysis, though these differences in difficulty are obviously worthy of further investigation, perhaps through more systematic use of verbal protocols. Thus it is impossible to say with any certainty whether our difficulty models worked because they captured common strategies used by all people or two different strategies both of which were regularly used. What we can conclude is that, at a general level, there is considerable consistency in the strategies used.

The success of our approach to predicting the difficulty of AR problems has practical applications. Although AR sets are no longer used in the GRE, they are still used in other tests, for example the Law School Admission Test (LSAT). What is more, meta-analysis of research into the GRE has shown that AR sets do correlate with performance in graduate school (Kuncel et al, 2001). Hence they do have some predictive validity and would seem to measure abilities which are important in determining future success. It is not possible to say what it is that makes the problems such good predictors, but one important factor might well be flexibility in strategy use.

The present research, which has involved both the development of a difficulty model and the writing of a computer program which can generate novel AR items, means that it is now possible to computer-generate AR items of predictable difficulty. Field testing by ETS suggests that the difficulty model is effective, with correlations of between .55 and .76 between predicted and actual difficulty. There is a very real possibility that, with a fairly small amount of further development, it will be possible to produce AR sets in real time, with novel items of predictable difficulty being automatically generated every time a candidate takes the test.

Where does this leave the theories of reasoning discussed in the introduction? Mental model theorists will take comfort from the finding that the semantic informativeness of a rule correlated negatively with difficulty (though we have also pointed out that this may be confounded with other factors), and from the finding that model variability score figured in our difficulty models. However, the various measures of numbers of possible orders did not figure in the difficulty models, suggesting that when people solve AR problems they do not routinely construct all possible orders. Rather, they try to construct partial models, and this is much easier the more fixed elements there are. Mental model theorists have often maintained that people work with partial models due to working memory constraints. However, the kind of schematic modelling we have observed is new: the partial mental models constructed contain the determinate parts of the problem, or those parts that are relatively easy to represent in a model.

The fact that the number and complexity of the rules needed to eliminate distracters or prove the key were important in predicting difficulty will resonate with advocates of mental logic, who claim that reasoning involves the application of internalised rules. In other words, people seem to use a rule-based eliminative strategy in which they apply rules individually and in combination to work out what conclusions follow and what conclusions can be eliminated. However, it is not clear that the use of a given set of rules, as happens in AR sets, is equivalent to the application of the kinds of internalised rules proposed by mental logicians. It has the appearance of a task-specific strategy rather than a general approach to reasoning.

The simple fact is that neither theory is able to provide a full explanation of our data. With complex problems such as those studied in this paper, it would seem that both approaches are used, though in our study they appear more like strategies than general theories (cf. Roberts, 1993). The theories have merit in explaining performance on certain relatively simple tasks, but in complex reasoning tasks of the kind used in this paper, neither theory is adequate on its own. The research agenda thus becomes one of identifying when each approach is used, not of deciding which theory is correct.

Manuscript received 22 September 2004
Revised manuscript received 14 February 2005

References

Bell, V. A., & Johnson-Laird, P. N. (1998). A model theory of modal reasoning. *Cognitive Science,* 22, 25–51.

Bradon, P., Evans, J. St. B. T., Handley, S., Newstead, S. E., Dennis, I., & Prat-Sala, M. (undated). *Development of algorithms for generating analytical reasoning problems. Report 4*. Unpublished report delivered to Educational Testing Service, Princeton.

Cohen, J. D., MacWhinney B., Flatt M., & Provost J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioural Research Methods, Instruments & Computers,* 25, 257–271.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472–517.

Educational Testing Service. (1996). *GRE. Practising to take the general test: Big Book*. Princeton, NJ: Educational Testing Service.

Evans, J. St. B. T., Handley, S. J., Harper, C. N. J., & Johnson-Laird, P. N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* 25, 1495–1513.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: the Psychology of deduction*. Hove, UK: Lawrence Erlbaum.

Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum.

Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1994). Why models rather than rules give a better account of propositional reasoning: A reply to Bonatti and to O'Brien, Braine, and Yang. *Psychological Review*, *101*, 734 – 739.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, *127*, 162 – 181.

Newstead, S. E., Griggs, R. A., & Chrostowski, J. J. (1984). Reasoning with realistic disjunctives. *Quarterly Journal of Experimental Psychology, 36A*, 611 – 627.

Potts, G. R. (1974). Storing and retrieving information about order relationships. *Journal of Experimental Psychology*, *103*, 431 – 439.

Rips, L. J. (1994). *The psychology of proof*. London: MIT Press.

Roberts, M. J. (1993). Human reasoning: Deduction rules or mental models, or both? *Quarterly Journal of Experimental Psychology, 46A*, 569 – 589.

Roberts, M. J., & Sykes, E. D. A. (2003). Belief bias and relational reasoning. *Quarterly Journal of Experimental Psychology, 56A*, 131 – 154.

Vandierendonck, A., & De Vooght, G. (1996). Evidence for mental-model-based reasoning: A comparison of reasoning with time and space concepts. *Thinking and Reasoning*, *2*, 249 – 272.

Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology*, *21*, 92 – 107.