Memory & Cognition
1998, 26 (5), 1089-1101

Individual differences in working memory and reasoning–remembering relationships in solving class-inclusion problems

MARK L. HOWE, F. MICHAEL RABINOWITZ, and T. LYNETTE POWELL
Memorial University of Newfoundland, St. John's, Newfoundland, Canada

In the present experiment, we evaluated the effects of individual differences in reading span and variation in memory demands on class-inclusion performance. One hundred twenty college students whose reading spans ranged from low to medium to high (as indexed by a computerized version of the Daneman and Carpenter [1980] reading-span task) solved 48 class-inclusion problems. Half of the subjects had the solution information available when the problems were presented; the other half performed a detection task between solution information and problem presentation. The results from both standard statistical analyses and from a mathematical model indicated that differences in reading span and memory load had predictable, similar effects. Specifically, the sophistication of reasoning strategies declined when memory demands increased or when reading spans decreased. Surprisingly, these effects were primarily additive. The results were interpreted in terms of global resource models and findings from the developmental literature.

In a recent series of experiments, we have used class-inclusion tasks to study reasoning–remembering tradeoffs in the development of human problem solving across the lifespan (Howe & Rabinowitz, 1996; Rabinowitz, Howe, & Lawrence, 1989). In general, the class-inclusion task involves the presentation of two subclasses: a major subclass (e.g., *there are six robins*) and a minor subclass (e.g., *there are four swallows*). Following this, subjects are asked an inclusion question involving the superordinate class (*birds*) and the major subclass (*Are there more robins or more birds?*). Additional questions can be asked involving the subclasses (*Are there more robins or more swallows?*) or the superordinate class and the minor subclass (*Are there more swallows or more birds?*).

When solving class-inclusion problems, subjects can use a variety of strategies. The most appropriate reasoning strategy reflects an understanding that the numerosity of the superordinate class has to be greater than or equal to the numerosity of the subclasses. This is known as class-inclusion reasoning. Alternatively, subjects can use an inappropriate strategy in which they treat the superordinate class as the subclass that is not specified in the problem. To illustrate, if the problem is a class-inclusion one of the form "Are there more robins or more birds?" and the subject treats "birds" as "swallows" (and remembers the relative numerosities), then the subject will answer the question incorrectly by responding "more robins." This is known as a subclass–subclass strategy and is the prevalent strategy in younger subjects (Howe & Rabinowitz, 1996). Interestingly, when task difficulty is increased, by increasing either memory or information demands, young adults in our studies also show subclass–subclass reasoning. That is, college students revert to a simpler form of reasoning, one that characterizes younger problem solvers.

In the present experiment, we were interested in whether we would obtain similar patterns as a function of individual differences in young adults' working memory. Specifically, we used Daneman and Carpenter's (1980) reading-span task as an index of working memory and predicted that adults with smaller reading spans would use more primitive solution strategies than would adults with larger reading spans. That is, because class-inclusion tasks can involve a tradeoff between remembering the premise (solution) information and implementing a reasoning strategy to solve the inclusion problem, subjects with fewer available resources tend to rely on simpler solution strategies that have fewer resource requirements.

We selected Daneman and Carpenter's (1980) measure because of the central role that it plays in the Just and Carpenter (1992) activation-based language comprehension model; because it satisfies many important criteria for measures of working memory (Howe & Rabinowitz, 1990); and because their conceptualization of working memory is similar to our earlier proposal (Rabinowitz et al., 1989). Specifically, this measure has been used successfully to predict individual differences in single-sentence comprehension tasks (see review by Just & Carpenter, 1992), including the resolution of syntactic ambiguity (Perlmutter & MacDonald, 1995), and it has been suggested

Preparation of this article was supported by Grants OGP0003334 (to M.L.H.) and OGP0002017 (to F.M.R.) from the Natural Science and Engineering Research Council of Canada. Correspondence concerning this article should be addressed to M. L. Howe, Department of Psychology, Memorial University of Newfoundland, St. John's, Newfoundland, A1B 3X9 Canada (e-mail: mhowe@morgan.ucs.mun.ca).

—**Accepted by previous associate editor Kathryn T. Spoehr**

as potentially relevant to more general problem-solving tasks (Cantor & Engle, 1993; Just & Carpenter, 1992; Waters & Caplan, 1996). Although this measure appears not to be systematically related to tasks that are predominantly automatic, such as obligatory language processing operations, it has been speculated that it is related to more consciously controlled processes, ones that occur in both language and, more generally, problem solving (Conway & Engle, 1994; Waters & Caplan, 1996). Indeed, the general importance of separating automatic and controlled aspects of performance in relation to resources has been emphasized for some time (see, e.g., Norman & Shallice, 1986).

We were also curious about whether reading span and manipulated memory load (e.g., solution information available simultaneously [no load] or not simultaneously [load] with the problem) would produce independent effects. Most resource theorists, especially those who assume a common pool of resources, would predict an interaction between working memory capacity, regardless of how it is defined, and task demands (see, e.g., for reviews, Howe & Rabinowitz, 1989, 1990; Just & Carpenter, 1992). That is, as task demands increase, deterioration of performance should be more evident in low than in high reading-span subjects. However, there do exist conditions under which independence would be obtained. To see one possible way in which additive effects could emerge, consider the following example. Suppose that the solution to a given problem involves the expenditure of resources on both remembering and reasoning. Assume that low reading-span subjects possess 20 resource (working memory) units and high reading-span subjects possess 40 units. Assume further that remembering (which precedes reasoning) costs 2 resource units in the no load conditions and 10 units in the load conditions. Finally, it must also be assumed that the probability that a sophisticated reasoning strategy can be used is linearly related to the number of remaining, unused resource units. Even with these assumptions, floor and ceiling effects would produce interactions. That is, reading-span scores below a low threshold would constrain subjects to the use of a primitive strategy, whereas scores above a high threshold would result in the use of a sophisticated strategy. However, in the class-inclusion task used here, it might be possible to observe additive effects, because we have not found floor or ceiling effects with college students (Howe & Rabinowitz, 1996; Rabinowitz et al., 1989). If additive effects were observed, then a number of functional constraints could be imposed on future modeling endeavors.

Our Class-Inclusion Problems

Our paradigm involves presenting a statement followed by a problem with three possible solutions. Each statement consisted of numerical information, color information, or both about items in two subclasses. The form of the statements when both types of information were presented was as follows: "There are *n*1 *c*1 *x*1s and *n*2 *c*2 *x*2s." The *n*s refer to number, the *c*s to color, and the *x*s to item type. An example of one of the statements is "There are 6 red robins and 4 brown swallows." One of 11 types of problems representing the combination of relevant dimension (number or color), number of items in the two subclasses (same or different; the color associated with each subclass was always different), and type of comparison (subclass–subclass, minor-subclass vs. class, class inclusion) followed each statement. It was impossible to present the 12th and missing minor-subclass versus class number-equal problem because there is no minor subclass unless the numbers of items in each subclass are different. When the numbers of items in the subclasses were the same, equivalence problems were used for both number and color dimensions; otherwise, superlative judgments were required. The form of the number-different problems was: "Are there more *x*1s or more *x*2s?" (e.g., "Are there more swallows or more birds?"). The form of the number-same problems was: "Are there the same number of *x*1s as *x*2s?" (e.g., "Are there the same number of robins as swallows?"). The form of the color-different problems was the following: "Is the *c*-est *x*1 *c*-er than the *c*-est *x*2?" (e.g., "Is the reddest swallow redder than the reddest bird?"). The form of the color-same problems was as follows: "Is the *c*-est *x*1 the same color as the *c*-est *x*2?" (e.g., "Is the reddest robin the same color as the reddest swallow?").

The same three alternative solutions, randomly ordered, followed each number-relevant problem: more *y*1s, more *y*2s, and same number. A different set of three alternatives, randomly ordered, followed each color-relevant problem: *y*1s *c*-er, *y*2s *c*-er, and same color. The *y*s stand for either the subclasses or the relevant superordinate class, and the *c*-ers stand for a superlative color label (e.g., *redder*). Thus, following the statement about robins and swallows, the number-relevant class-inclusion choices would have been *more robins*, *more birds*, and *same number*. The color-relevant subclass–subclass choices would have been *robins redder*, *swallows redder*, and *same color*. The dependent variable of interest was the number of correct choices.

We have acknowledged in our previous articles (Howe & Rabinowitz, 1996; Rabinowitz et al., 1989) that color problems represent a variant on the traditional class-inclusion problems, one that increases both syntactic and semantic complexity. However, the advantage of using such problems is that it allows the study of class-inclusion reasoning in a domain that necessitates memory. Whereas number-based class-inclusion problems can be answered on purely logical grounds, color-based problems demand that subjects remember the premises. Our previous research has shown that this advantage outweighs the changes in complexity associated with color problems and that the latter changes do not represent a fundamental alteration of the classic class-inclusion problem.

Understanding Reasoning–Remembering Relationships in Class-Inclusion Tasks

We have developed a simple model that characterizes the manner in which groups of subjects represent the information in the statements and interpret the problems when

choosing among the three possible solutions to each problem. The model contains five parameters: two that are related to memory (e and d) and three, to reasoning (i, s, and u). The two memory-related parameters in the model are estimates of the manner in which information is encoded at retrieval (i.e., when problems are solved). Here e gives the probability of correctly encoding the values on the relevant dimension as same or different on the test, while $1 - e$ represents the wrong encoding of the relevant dimension as same or different. The parameter d, which is irrelevant to the number-same problems, represents the conditional probability of remembering the cue values associated with each subclass when the cues are correctly encoded as different. The conditional probability of associating each of the two cue values with the wrong subclass is $1 - d$. Although these estimates are influenced by the way in which information is stored when the statements are presented and by the loss between initial presentation and the presentation of the problem, they do not reflect these processes directly. Three parameters are used to estimate the way problems involving the superordinate class and either subclass are interpreted (i = idiosyncratic, s = subclass–subclass, u = understanding; see also Hodkin, 1987; Howe & Rabinowitz, 1996; Rabinowitz et al., 1989). Since these interpretations are treated as mutually exclusive and exhaustive in the model, only two degrees of freedom are lost in their estimation:

$$1 = i + s + u. \quad (1)$$

Thus, in the present study, four free parameters were estimated for the number-different and color problems, but only three free parameters were estimated for the number-same problems. The parameters and associated definitions are summarized in Table 1 and are operationalized in the equations that appear in Appendix B.

Different sets of equations (see the following sections and Appendix B) were constructed for the number-same and number-different problems, but one set of equations was sufficient for the color-relevant problems because different colors were always associated with each of the two subclasses appearing in a statement. Note that abbreviations follow the equation numbers in order to specify the appropriate reference(s) for each equation: number different (*n*d), number same (*n*s), and color (*c*). It was assumed that subjects always interpreted subclass–subclass problems correctly.

Table 1
Theoretical Definitions of the Choice Model's Parameters

Parameter	Theoretical Definition
Memory	
e	Probability of correct encoding of dimensional cues as same or different.
d	Probability of correctly associating the cue with each of the two subclasses given that the relevant dimension is accurately encoded as different.
Reasoning	
u	Probability of understanding and accurately interpreting questions involving comparison of the superordinate class and subclass.
s	Probability of subclass-subclass interpretations of questions involving comparisons of the superordinate class and a subclass.
i	Probability of idiosyncratic interpretations of questions involving the superordinate class and a subclass.

Specific Predictions

In using this task with college-aged subjects, we have consistently found that they switch from class-inclusion reasoning to subclass–subclass reasoning when memory load increases. We expected to replicate this finding. Similarly, if the reading span measure would generalize to a broader problem-solving context so that resource differences would be reflected in individual differences in reading span, then we should see a similar switch from inclusion reasoning to subclass–subclass reasoning as reading span decreased. Finally, it was of interest to determine whether reading span and memory load effects might be additive. The memory parameters estimated in our prior research suggested that college students were not operating at ceiling in the no load conditions or at floor in the load conditions (Howe & Rabinowitz, 1996; Rabinowitz et al., 1989). If the use of reasoning strategies is actually linearly related to resource availability, additive effects would occur.

METHOD

Subjects

The 60 males and 60 females were university students who were paid for their participation.

Apparatus and Materials

Two different computer programs were developed for this experiment, one for the reading-span task and one for the class-inclusion task. The reading-span task was programmed in Quick Basic and the class-inclusion task was programmed in Turbo Basic. Stimuli were presented on a 22.5 × 27 cm monitor, and the subjects entered their responses on a 101-key enhanced keyboard.

The reading-span test was constructed from 110 English sentences ranging from 10 to 20 words in length. The sentences were taken from books of general knowledge. Half of the sentences were true and half of them were false. They were arranged in sets containing different numbers of sentences (2, 3, 4, 5, or 6). There were five sets of each type (e.g., five sets containing 2 sentences, five sets containing 3 sentences). For each subject, the sentences were randomly assigned to sets. After each sentence was presented, the subject was required to verify whether it was true or false, and, following presentation of all the sentences in a set, to free recall the final word appearing in each sentence.

In the class-inclusion task (see Rabinowitz et al., 1989, Experiment 2), the computer was used to control presentation of verbal materials and to record choices. The subjects responded by pressing the numerals 1, 2, and 3 on the number pad. The material was arranged in blocks of 48 units (see Appendix A), with each unit comprising a statement (e.g., "There are 6 red robins and 4 brown swallows"), a problem (e.g., "Are there more robins or more birds?"), and three alternatives to choose from (e.g., more robins, more birds, same number). As in the examples, each statement consisted of two numbers (n_1 and n_2), two colors (c_1 and c_2), and two nouns. The values of the

numbers ranged from 4 to 9, where $n_1 = n_2$ for the equivalence problems and $n_1 \neq n_2$ for the nonequivalence problems.

Each of the 11 possible problem types determined by question, numerosity, and dimension, plus an additional number-different minor-subclass versus class problem, appeared once in successive blocks of 12 problems. The additional number-different minor-subclass versus class problems were substituted for the impossible-to-construct number-same minor-subclass versus class problems. For each individual, the numbers were randomly assigned to each problem and the problems were quasi-randomly assigned, subject to the blocking constraint.

Design

A 3 (working memory: low vs. medium vs. high reading span) × 2 (processing load: no load vs. detect) × 2 (gender) factorial design was used for testing. All subjects were first tested for working memory capacity by using the reading-span test (also see Daneman & Carpenter, 1980; Daneman & Green, 1986; Just & Carpenter, 1992). On the basis of their scores, the subjects were divided into low, medium, and high reading-span groups. The subjects were grouped by reading span and gender. Group members were then randomly assigned to one of two processing load conditions—namely, the *no load* and *detect* conditions.

Procedure

All subjects were tested individually. For the reading-span task, instructions appeared on the monitor and any questions were answered by the experimenter. This was followed by a pretest, which served to familiarize the subjects with the procedure and was discontinued when an individual correctly recalled the final words in one two-sentence set or when a total of five two-sentence sets had been shown. The subjects were then presented sets of sentences, with the number of sentences per set increasing across trials. A trial consisted of five sets of sentences, with each set comprising a fixed number of sentences. All subjects began with the two-sentence test. They were informed how many sentences to expect, and each sentence in a set would then appear on the screen individually for 8 sec. After presentation of a sentence, the subjects had 5 sec to verify whether the statement was true or false by pressing "t" or "f" on the keyboard. Note that the "true" and "false" responses were used as a control to ensure that the subjects read and processed each entire sentence rather than simply memorized the last word in each sentence. Following verification, the next sentence in the set was presented. At the end of each set, the subject was asked to recall by typing, in any order, the final word from each of the sentences in that set. Recall was self-paced, and the subjects indicated unrecallable words by hitting the carriage return. Only the first three letters of the words that the subjects entered were evaluated by the computer. This was done in order to reduce the typing and spelling demands on the subjects. This process was repeated for five sets of n sentences, where n ranged from two to six. If subjects successfully recalled the words in three of five sets, they would proceed to the next level, in which the number of sentences was increased by one. Three seconds elapsed between the end of a recall task and the beginning of the next sentence set.

Reading-span scores could range from 1 to 6. If subjects were successful in only two of the five sets, a reading span was assigned that was midway between that particular set number and the previous set number. For example, if a subject was successful in only two sets in the three-sentence group, a reading-span score of 2.5 would be assigned. If a subject's recall score was less than two for a set, that subject would be assigned a reading-span score corresponding to that for the previous set (e.g., 2 in the preceding, three-sentence example). Upon completion of the reading-span task, the subjects were divided into three groups on the basis of their assigned reading-span scores. A score of 4.0 or higher was considered *high span*, a score of 3.0 to 3.5 was considered *medium span*, and a score of less than 3.0 was considered *low span*.

Following a delay equivalent to the time that it took to load the software, each subject was next tested in the class-inclusion phase of the experiment. The instructions appeared on the monitor and contained the following information: There would be 48 problems, each problem would be preceded by a statement containing the information needed to solve the problem, and the problems should be solved by pushing the correct button on the number pad of the keyboard. Any questions were answered by the experimenter, and the subject then pressed any key to start the experiment.

In the no load condition, a trial consisted of the presentation of a statement that remained on the monitor until the end of the trial. After reading the statement, the subject pressed any key to initiate the presentation of a problem and related choices that also remained on the screen until a response was made. There was a 1-sec delay between the end of one trial and the beginning of the next.[1]

The instructions also appeared on the screen for the detect condition. Here, a trial consisted of the presentation of a statement, followed by a letter-detection task, followed by a problem. In the detection task, subjects were to press "1" if a "y" appeared on the left side of the screen, "2" if a "y" appeared on the right side of the screen, or a "3" if a "y" appeared on both the left and right sides of the screen. Consistent with the values of up to five random variables, the "y" could appear anywhere on the left, right, or both sides of the screen. The timing of events for the detect condition was as follows. First, a statement was presented, and it remained on the screen until a key was pressed by the subject. Then a blank screen appeared for 1 sec, followed by a READY signal, which appeared for 100 msec in the center of the monitor with the letters arranged vertically to bisect the screen into left and right halves. The screen then went blank for 500 msec and a "y" then appeared on the screen for 20 msec, followed immediately by "1 = left," "2 = right," and "3 = both." If there was no response within 2 sec, then "You took more than 2 seconds. Please try to respond within two seconds." appeared on the screen for 2 sec. Following either a legitimate response or the offset of the *respond within 2 sec* reminder, the screen went blank for 1 sec before the problem, and alternatives appeared. After the subject had solved the problem, the screen went blank for 1 sec before the next statement appeared. Thus, for subjects in the detect condition, the time between the offset of the statement and the onset of the problem varied between 2.62 (plus detection-response latency) and 6.62 sec. Note that the no load and detect conditions differed not only in the presence of a secondary task, but also in the availability of statement information when questions were answered and in the delay between statement and question presentation.

RESULTS

Because the descriptions generated with the mathematical models were of primary interest, analysis of variance (ANOVA) of the choice data will be described selectively. Furthermore, the statistical tests associated with the ANOVA were powerful; the 120 subjects each generated 48 data points. A .01 significance level was adopted for the between-subjects effects and a .001 significance level was adopted for the within-subjects effects in order to reduce the probability of Type I errors in the ANOVA. The analyses that bear directly on the mathematical models will be reported at the .05 level, because these tests involved fewer degrees of freedom and often could be used to reject the model when effects were significant. In the ANOVA, reading level (low, medium, or high), gender

(female or male), and memory load (no load or detect) were the between-subjects variables. The within-subjects variables were 4 blocks of 12 trials (Trials 1–12, 13–24, 25–36, and 37–48), numerosity in the statements (different or same), dimension (number or color), and problem type (subclass–subclass, minor subclass–class, class inclusion).

Before we present the outcome of this ANOVA, we will report some findings concerning our individual difference index of working memory—namely, reading span. In a series of regression analyses in which gender and reading span were the predictor variables, a significant proportion of the variance was accounted for in the percentage correct on the true–false questions [$F(2,118) = 5.10$, $p < .008$, $R^2 = .08$], latencies for answering the true–false questions [$F(2,118) = 9.98$, $p < .0001$, $R^2 = .14$], and recall latency [$F(2,118) = 14.59$, $p < .00001$, $R^2 = .20$]. The partial correlational analyses revealed that only reading span was a significant predictor for each of these variables. The simple correlations between reading span and the dependent measures were .28, −.36, and −.43, respectively. As we will report subsequently, the reading-span measure was reliable in the ANOVA and had dramatic consequences on the parameters of our mathematical choice model.

Analysis of Variance of the Choice Data

For the ANOVA, each correct response was scored "2," errors consistent with appropriate same–different encoding of the dimension relevant to the problem (E_1 errors) were scored "1," and the remaining errors (E_2 errors) were scored "0."[2] For example, the E_1 error associated with the number-different class-inclusion problem involving robins and birds would be "more robins," whereas the E_2 error would be "same number." The E_1 errors associated with the number-different minor subclass–class and subclass–subclass problems would be "more swallows"; the E_2 errors would be "same number." The E_1 error associated with the color-different class-inclusion problem involving robins and birds would be "robins redder"; the E_2 error would be "birds redder." It should be emphasized that the E_2 class-inclusion color error is different from all the other types of E_2 errors because it *cannot* follow same–different encoding errors (i.e., same color encoding) unless the encoding error is associated with idiosyncratic interpretation. The E_1 error associated with the color-different minor subclass–class and subclass–subclass problems would be "swallows redder"; the E_2 error would be "same color."

To begin, performance varied linearly as a function of working memory [$F(2,109) = 5.27$], although differences were only statistically reliable between the high ($M = 1.70$) and low ($M = 1.53$) reading-span subjects, with medium reading-span subjects falling somewhere in between ($M = 1.61$). Because the more sensitive mathematical model analyses showed reliable differences among all three groups, data for the medium reading-span group were retained rather than eliminated, as in previous studies (for a review, see Just & Carpenter, 1992). There was a small improvement in performance across blocks [Ms = 1.53, 1.61, 1.64, 1.67; $F(3,327) = 15.85$]. Because the remaining significant effects were lower order to the three-way interactions of memory load × numerosity × problem type [$F(2,218) = 7.31$] (see Figure 1) and dimension × problem type × numerosity [$F(2,218) = 14.08$] (see Figure 2), they replicated our previous findings (Howe & Rabinowitz, 1996; Rabinowitz et al., 1989), and none involved the variable of primary interest in this article, reading span, they will be discussed only in summary form. As can be seen in Figure 1, when there was no memory load, subclass–subclass problems were answered best, class-inclusion problems next, and smaller subclass–class problems poorest, regardless of numerosity. However, in the presence of an additional memory load, both subclass–subclass and smaller subclass–class problems were answered better than class-inclusion problems, with no difference between the former two when numerosity differed. This figure also shows that the main effect of memory load is interpretable. That is, subjects were correct more often when there was no memory load ($M = 1.71$) than when there was [$M = 1.52$; $F(1,109) = 19.72$]. As can be seen in Figure 2, number problems were answered better than color problems with subclass–subclass and smaller subclass–class questions, but not with class-inclusion questions, particularly when numerosity was the same. Analyses based on the mathematical model reveal that class-inclusion performance was facilitated on the color problems because subjects tended to make more encoding errors on color than on number problems. A consequence of erroneously encoding the colors as same is that color subclass–subclass and smaller subclass–class questions are answered incorrectly while class-inclusion questions are answered correctly when either class-inclusion or subclass–subclass reasoning strategies are used.

Modeling Class-Inclusion Choice Data

Before we can use the mathematical model to interpret reading-span differences in reasoning and remembering, the parameters must be estimated and the degree of fit of the model to the choice data evaluated statistically. In order to see how this process works, we will begin with how we define the data space. For each reading span, dimension (color or number), numerosity of the subclasses (same or different), and memory load (no load or detect), three data types were defined for each question type (subclass–subclass, minor subclass vs. class, class inclusion): S = success, E_1 = error associated with correctly encoding the values on each subclass as same or different, and E_2 = remaining error usually associated with incorrectly encoding the values on each subclass as same or different. Different equations were developed for the subclass–subclass (ss), minor-subclass versus class (msc), and major-subclass versus class or class-inclusion (ci) questions. These equations and the associated likelihood functions appear in Appendix B.

Because

$$1 = P(S) + P(E_1) + P(E_2), \qquad (2, nd, c, ns)$$

two degrees of freedom are associated with the data for each question type. Thus, when three question types are available to test the model, as in the color problems, a total of six degrees of freedom exist in the data. Two degrees of freedom are necessary to estimate the memory parameters (e and d) and two more are needed to estimate the reasoning parameters (see Equation 1). This leaves two degrees of freedom for assessing the goodness of fit of the model for the color problems. Similarly, because there were no minor-subclass versus class number-same problems, the three number-same parameters were estimated using a data set containing four degrees of freedom. Finally, because number-different minor-subclass versus class problems were used in lieu of the impossible minor-subclass versus class number-same problems, the four number-different parameters were estimated using a data set containing eight degrees of freedom (the two sets of minor-subclass vs. class problems were treated independently).

Two different maximum-likelihood parameter estimation procedures were used. The first was a stochastic algorithm (Rabinowitz, 1995), and the second involved a simplex method (Siddal & Bonham, 1974). These procedures yielded identical parameter estimates (one strong indication that there was good agreement between the model and the data).

The model, the data space, and the estimation procedure having been defined, only two steps remain: direct assessment of goodness of fit and hypothesis testing. Concerning goodness of fit, two tests are conducted, a necessity and a sufficiency test. The necessity test examines whether a model with fewer parameters provides a statistically adequate account of the data. This involves comparisons of the likelihood of the data given three- (number-same questions) or four-parameter models with the likelihood of the data given a simpler two-parameter model, one in which memory is assumed to be perfect (i.e., $1 = e = d$). Specifically, the necessity test evaluates the null hypothesis that a model with fewer parameters fits the data as well as a model with more parameters. Rejection of this null hypothesis means, in this case, that the model with three (number-same problems) or four parameters is necessary to account for this data and that memory encoding was not perfect. Thus, this test has a dual purpose. First, it serves to evaluate the necessity of having more (three or four) rather than less (two) parameters in the model, and second, it serves to confirm the necessity of positing memory processes to account for class-inclusion reasoning. The necessity tests for each reading span (by condition and question) can be found in Table 2. As can be seen, in the majority of cases the null

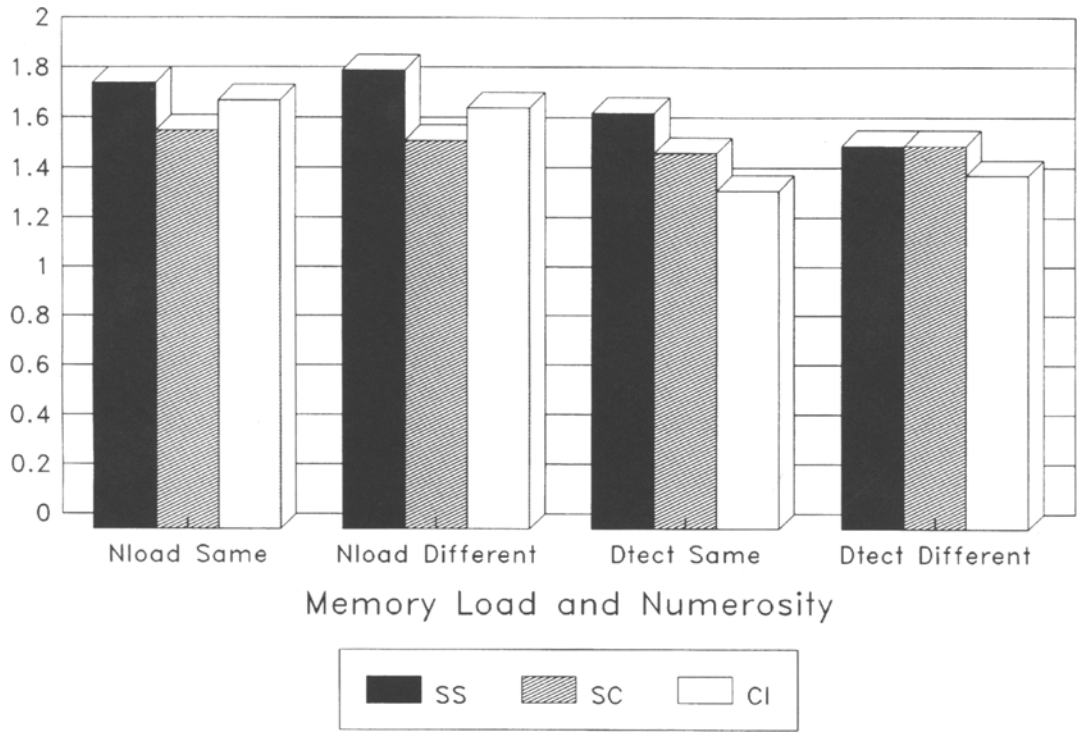

Figure 1. Means for the memory load × numerosity × problem type interaction. Nload = no load; Dtect = detect; SS = subclass–subclass; SC = minor subclass–class; CI = class inclusion.

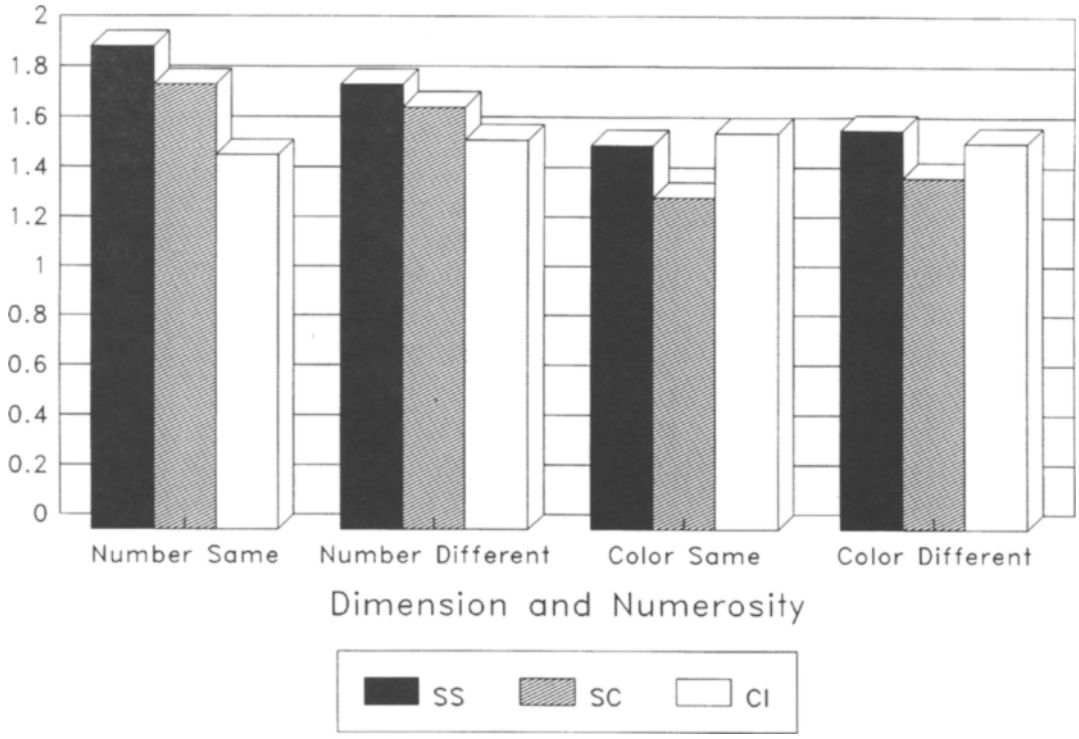

Figure 2. Means for the dimension × numerosity × problem type interaction. SS = subclass–subclass; SC = minor subclass–class; CI = class inclusion.

hypothesis can be soundly rejected. Exceptions occurred for (1) all number problems in the no load condition and (2) the number-same problems for the medium and high reading-span groups in the detect condition. In "easy" conditions, in which memory demands were low, or in groups with higher reading-span scores, memory was at ceiling. This former result has been obtained previously (e.g., Rabinowitz et al., 1989).

The sufficiency test, which examines whether a more complex model is required to account for the data, involves comparing the likelihood of the data given the three- (number-same problems) or four-parameter models with the likelihood of the data itself (i.e., when all the empirical probabilities are free to vary, thus exhausting all of the information in the data). What this means is that the theoretical model (which reduces the number of parameters estimated) is compared with a data-based model (which does not limit the number of parameters estimated). Specifically, the sufficiency test, like the necessity tests, evaluates the null hypothesis that a model with fewer parameters fit the data as well as a model with more parameters. Failure to reject the null hypothesis means, in this case, that the model with three (number-same questions) or four parameters is sufficient to account for the data. The sufficiency tests for each reading span by condition combination can also be found in Table 2. As can be seen, in the majority of cases the null hypothesis could not be rejected. Exceptions occurred for (1) color-same problems in the no load condition for high reading-span subjects (significant at $p < .01$, but not at .001) and (2) number-different problems in the detect condition for the low and medium reading-span subjects (neither of the χ^2 tests were significant at $p < .001$).

Despite these exceptions in both the necessity and sufficiency tests, the three- (number-same questions) and four-parameter models were found to be, on the average, both necessary (i.e., more than two parameters were needed to account for much of the data) and sufficient (i.e., generally no more than three or four parameters were needed to account for the data). Note that when the latter conclusion did not hold at $p < .01$ (but did at $p < .001$), the magnitude of the corresponding necessity tests was substantially greater than that of the sufficiency tests, suggesting that the model was accounting for a considerable proportion of the variance in the data (even though it was not as adequate as the data itself). Thus, although in some cases the data may have been somewhat more complicated than our models depicted, it can be reasonably concluded that, within the bounds of statistical tolerance, the three- (number-same problem) and four-parameter models provided an adequate and parsimonious fit to the data from this experiment.

Next, we can turn our attention to the main business of hypothesis testing. Because the parameter estimates from the fitted models are identifiable, they can be used directly in testing hypotheses about the theoretical rela-

Table 2
Goodness-of-Fit Tests

Condition	Necessity Test	Sufficiency Test
	No Load	
Low Reading Span		
Number different	$\chi^2(2) = 5.55$	$\chi^2(4) = 10.48$
Number same	$\chi^2(1) = 0.01$	$\chi^2(1) = 1.80$
Color different	$\chi^2(2) = 173.50$*	$\chi^2(2) = 3.98$
Color same	$\chi^2(2) = 216.02$*	$\chi^2(2) = 8.48$
Medium Reading Span		
Number different	$\chi^2(2) = 3.82$	$\chi^2(4) = 5.94$
Number same	$\chi^2(1) = 1.22$	$\chi^2(1) = 1.39$
Color different	$\chi^2(2) = 100.42$*	$\chi^2(2) = 5.14$
Color same	$\chi^2(2) = 144.34$*	$\chi^2(2) = 2.99$
High Reading Span		
Number different	$\chi^2(2) = 3.10$	$\chi^2(4) = 12.56$
Number same	$\chi^2(1) = 1.73$	$\chi^2(1) = 0.00$
Color different	$\chi^2(2) = 13.71$*	$\chi^2(2) = 8.43$
Color same	$\chi^2(2) = 38.24$*	$\chi^2(2) = 11.50$*
	Detect	
Low Reading Span		
Number different	$\chi^2(2) = 170.79$*	$\chi^2(4) = 17.51$*
Number same	$\chi^2(1) = 17.99$*	$\chi^2(1) = 1.33$
Color different	$\chi^2(2) = 234.26$*	$\chi^2(2) = 5.75$
Color same	$\chi^2(2) = 235.49$*	$\chi^2(2) = 0.09$
Medium Reading Span		
Number different	$\chi^2(2) = 232.16$*	$\chi^2(4) = 17.00$*
Number same	$\chi^2(1) = 0.14$	$\chi^2(1) = 1.39$
Color different	$\chi^2(2) = 92.45$	$\chi^2(2) = 2.97$
Color same	$\chi^2(2) = 205.33$*	$\chi^2(2) = 4.09$
High Reading Span		
Number different	$\chi^2(2) = 208.65$*	$\chi^2(4) = 8.07$
Number same	$\chi^2(1) = 0.42$	$\chi^2(1) = 0.11$
Color different	$\chi^2(2) = 117.70$*	$\chi^2(2) = 3.22$
Color same	$\chi^2(2) = 107.72$*	$\chi^2(2) = 5.51$

*$p < .01$.

tionships between reasoning and remembering and individual differences in reading span. The models' parameters are identifiable because (1) there were more data points than parameters, (2) both fit procedures gave the same parameter estimates, and (3) the parameter estimates generated using the maximum likelihood procedure were independent of the initial starting values. The numerical values of these parameters are given in Table 3.

The three-phase hypothesis-testing sequence begins with an experimentwise test that, like an omnibus F test, evaluates the null hypothesis that, on the average, the numerical estimates of the model's parameters did not vary statistically across reading spans and conditions (memory load). In each case, the null hypothesis was rejected, with the numerical values as follows: $\chi^2(20) = 222.93$, $p < .001$ (number-different problem); $\chi^2(15) = 102.41$, $p < .001$ (number-same problem); $\chi^2(20) = 111.95$, $p < .001$ (color-different problem); $\chi^2(20) = 111.86$, $p < .001$ (color-same problem).

The next phase involves conditionwise tests which, like t tests, evaluate the null hypothesis that the numerical estimates of the model's parameters do not vary statistically between pairs of conditions. Because in the present experiment there was a total of 36 conditionwise tests, the numerical results are given in Table 4. Concerning memory-load effects, it can be seen from Table 4 that the no load versus detect manipulation affected all reading-span groups (except the medium span group for the color-different problem), but to varying degrees. Concerning individual differences, 7 of the 12 comparisons were significant within the no load condition, and 9 of the 12 comparisons were significant within the detect condition.

Finally, parameterwise tests are conducted to determine the locus (reasoning and/or remembering) of these differences. Here, for the pairs of conditions that differed significantly, a series of $\chi^2(1)$ tests were conducted to determine which of the parameters differed reliably between the conditions. Because these tests are both tedious and space consuming to report, they are typically described in summary form. Consistent with this tradition, we present only the parameterwise differences that were statistically reliable ($p < .05$).

Memory-load effects. The effects of manipulating memory load on the model's parameters tended not to depend on either dimension or numerosity (i.e., number different, color different, color same) or on the reading-span level of the subject. That is, the memory-load manip-

Table 3
Estimated Parameter Values for the Choice Model

Condition	Parameter e	d	i	s	u
	No Load				
Low Reading Span					
Number different	1.00	1.00	.18	.26	.55
Number same	.91		.00	.46	.54
Color different	.74	.99	.07	.37	.56
Color same	.72	.97	.03	.34	.63
Medium Reading Span					
Number different	.97	.97	.26	.06	.68
Number same	.99		.04	.29	.68
Color different	.78	.96	.07	.24	.69
Color same	.72	.95	.08	.20	.73
High Reading Span					
Number different	1.00	.97	.14	.08	.78
Number same	1.00		.04	.15	.81
Color different	.93	.98	.15	.18	.67
Color same	.88	.99	.08	.18	.74
	Detect				
Low Reading Span					
Number different	.92	.73	.15	.69	.16
Number same	.92		.31	.54	.15
Color different	.75	.81	.20	.72	.08
Color same	.72	.82	.23	.57	.20
Medium Reading Span					
Number different	.96	.72	.00	.78	.22
Number same	.99		.12	.53	.35
Color different	.84	.92	.06	.41	.53
Color same	.81	.83	.00	.39	.61
High Reading Span					
Number different	.88	.68	.06	.40	.54
Number same	.89		.29	.48	.49
Color different	.84	.90	.07	.44	.49
Color same	.82	.90	.20	.29	.52

Table 4
Conditionwise Tests

Hypothesis	Number Different $\chi^2(4)$	Number Same $\chi^2(3)$	Color Different $\chi^2(4)$	Color Same $\chi^2(4)$
Memory Load (No Load vs. Detect)				
Low reading span	73.38‡	30.90‡	53.50‡	45.63‡
Medium reading span	59.12‡	12.20†	5.95	12.99*
High reading span	53.79‡	30.36‡	16.36†	23.32‡
Individual Differences (Low vs. Medium vs. High Reading spans)				
Within No Load				
Low vs. medium	14.24†	5.47	3.20	3.08
Medium vs. high	9.25	5.19	10.43*	14.80†
Low vs. high	14.98†	19.60‡	18.38†	14.50†
Within Detect				
Low vs. medium	8.89	8.77*	26.61‡	12.62*
Medium vs. high	19.69‡	8.56*	0.90	2.97
Low vs. high	23.68‡	13.99†	21.87‡	13.71†

*$p < .05$. †$p < .01$. ‡$p < .001$.

ulation uniformly affected the memory parameter *d* (in all but the number-same conditions, of course) by decreasing its value in the detect as compared with the no load condition. What this indicates is that low, medium, and high reading-span subjects were more likely in the detect than in the no load conditions to associate the wrong cue value (e.g., thought robins were brown instead of red) with a dimension (e.g., color), which is consistent with the assumption that specific information is more difficult to remember under high than under low memory-load conditions. Not only were these qualitative patterns relatively constant across reading span, so too were the quantitative patterns. That is, the average numerical differences in the parameter *d* favoring the no load over the detect conditions were relatively similar across low (.20), medium (.13), and high (.12) reading-span subjects. The parameter *e* was also affected in a similar manner, but only for the high reading-span subjects on the number-same and color-different problems. What this suggests is that although the effects of memory load and individual differences in reading span on the parameter *d* were additive, such was not the case with the parameter *e*.

The memory-load manipulation not only affected memory but, consistent with our previous work (Howe & Rabinowitz, 1996; Rabinowitz et al., 1989), also affected reasoning. In particular, regardless of problem type and reading-span level, the parameter *u* was smaller (all comparisons) and the parameters *s* (all comparisons except low and high reading-span subjects on the number-same problem) and *i* (all comparisons except low reading-span subjects on the number-different problem; medium reading-span subjects on both color problems and the number-same problems; and high reading-span subjects on the color-different problems) were larger in the detect than in the no load conditions. What this indicates is that subjects were less likely to use class-inclusion reasoning and more likely to use subclass–subclass reasoning or idiosyncratic responding and guessing in the detect than in the no load conditions. Thus, as in our other research, changes in memory load affected both memory and reasoning parameters, suggesting that a common resource was used in solving class-inclusion problems.[3]

Not only were these qualitative patterns relatively constant across reading span, so too were the quantitative patterns for the number problems. Specifically, regardless of whether subjects were in the low, medium, or high reading-span groups, the average numerical difference in the parameter *u* favoring the no load over the detect conditions (.39, .40, and .28, respectively) and the parameter *s* favoring the detect over the no load condition (.26, .48, and .33, respectively) were relatively constant in the number problems, although the difference in the *s* parameter was somewhat larger in the medium reading-span group. It would seem, therefore, that consistent with the ANOVA, the effects of memory load and individual differences in reading span were independent for number problems. However, the effects of memory load were more pronounced in the low than in the medium or high reading-span groups for the color problems. Specifically, the average numerical difference in the parameter *u* favoring the no load over the detect conditions was larger in the low (.44) than in the medium (.14) or high (.20) reading-span groups. Similarly, the average numerical value of the parameter *s* favoring the detect over the no load condition was larger in the low (.29) than in the medium (.18) or high (.19) reading-span groups. Thus, although the direction of the effects of memory load and individual differences in reading span on reasoning were the same for both number and color problems, the magnitudes of these effects did differ on color problems.

Individual differences. The trends here are also rather straightforward. In conditions where memory demands were low (the no load conditions), differences on number problems were confined to reasoning parameters. That is, for number problems, the higher a subject's reading span, the more likely they were to use class-inclusion reasoning and the less likely they were to use subclass–subclass

reasoning. In particular, the parameter *u* was higher and the parameter *s* lower in medium than in low reading-span groups and in high than in medium reading-span groups. The same pattern held for color problems, but there was an additional memory effect. Specifically, the parameter *e* tended to be larger in medium than in low and in high than in medium reading-span groups. This indicates that for the more difficult color problems, subjects with higher reading spans not only tended to use class-inclusion reasoning but also were better able to encode whether the cues were the same or different. Finally, when memory demands were increased (the detect conditions) these patterns continued. That is, *u* increased and *s* decreased with increasing levels of reading span. In addition, the parameter *i* tended to be higher in lower reading-span groups, and the memory parameters *e* and *d* tended to be smaller.

Interestingly, an examination of the quantitative patterns associated with the average numerical difference in the reasoning parameters across the different reading spans showed that larger differences existed between the medium- and low-span subjects than between the high- and medium-span subjects. Specifically, the average difference in *u* favoring the higher levels of reading span were greater between the low- and medium-span subjects (.13 in the no load condition and .28 in the detect condition) than between the medium- and high-span subjects (.06 in the no load condition and .08 in the detect condition). The average numerical difference in *s* favoring lower levels of reading span were in the same direction for the no load condition (low vs. medium = .10; medium vs. high = .05). The direction was reversed in the detect condition (low vs. medium = .10; medium vs. high = .13).

What these results indicate is that the higher the subject's reading span, the better the subject's memory and reasoning was in solving class-inclusion problems. Furthermore, reading-span effects were observed in both the memory and the reasoning parameters. It would seem, therefore, that this reading-span measure gives a good index of a common resource, one that can be deployed to support memory and reasoning functions in problem-solving tasks (see note 3).

DISCUSSION

The results of this study are straightforward and in line with our predictions. First, the main effect of manipulating memory load was on reasoning parameters. Consistent with our previous research, we found that college students reverted to subclass–subclass reasoning as memory load increased. Interestingly, we also found some small but reliable changes in the memory parameters favoring the no load condition. In a previous study (Rabinowitz et al., 1989) in which we used the same no load and detect conditions, we found a similar pattern with the memory parameters, but only two of seven comparisons were significant. We suspect that grouping subjects by reading-span levels reduced the variability of the parameter estimates in the present study, yielding more sensitive statistical tests. In general, the pattern of findings obtained here with the memory load manipulation is compatible with models in which reasoning and remembering trade off in the service of problem solving.

Second, as predicted, individual differences in reading span were consistently related to reasoning strategy selection for both the no load and the detect conditions. Furthermore, differences were obtained across each of the three reading-span levels (low, medium, and high), emphasizing the combined sensitivities of the model and the individual difference measure. In addition, in the detect condition, reading span affected the memory parameters, whereas in the no load condition, memory effects were obtained only for the more difficult color problems. In general, then, the effects of the individual difference manipulation and the memory-load manipulation were similar and are consistent with a single, global resource model (e.g., Norman & Shallice, 1986).

These findings confirm a number of suggestions that the reading span measure would be related to performance on problem-solving tasks (Cantor & Engle, 1993; Just & Carpenter, 1992; Waters & Caplan, 1996). Interestingly, reliable differences were obtained both between each of the reading-span levels and within each reading-span level as a function of memory load. This indicates that at least in this context, it is important to consider the entire reading-span scale and not just subjects representing the extremes of the scale. Recall that in the language comprehension studies in which this measure has been used most frequently, reported differences have been confined to comparisons involving low- versus high-span subjects (Just & Carpenter, 1992). Whether the increased sensitivity observed here is due to the use of a mathematical model, the extension of the reading span measure to problem-solving situations, or both remains an empirical question.

Third, and arguably surprising, the effects of memory load and individual differences in reading span were primarily additive. Although there were some differences in the magnitude, but not direction, of parameter differences across color and number problems, the overwhelming tendency was for memory load and individual differences in reading span to make independent contributions to performance. Apparently the special conditions that produce additive effects, described earlier, hold at least in part for college students. That is, subjects' reasoning performance was above floor and below ceiling with all combinations of memory load and reading span studied. It is very unlikely that these results would be obtained with younger or older subjects, because they typically use the more primitive subclass–subclass reasoning strategy under all conditions that we have examined (see Howe & Rabinowitz, 1996). Moreover, the present findings are consistent with the assumption that a quasi-linear relationship

exists between resources and strategy selection, at least in the restricted resource range sampled here. What this implies is that mathematical functions relating individual differences in working memory and resource-dependent performance are likely to be monotonic.

Because of the clear and compelling nature of our findings, it would seem worthwhile to extend Just and Carpenter's (1992) computer simulation to domains other than language comprehension. Indeed, consistent with a number of suggestions, reading-span measures are predictive of performance in problem-solving situations (e.g., Cantor & Engle, 1993; Howe & Rabinowitz, 1996; Waters & Caplan, 1996). The present findings also add to the growing database favoring the existence of a single, common resource. Because our class-inclusion problems were presented linguistically, our data do not rule out the possibility that we are dealing with a single language-based resource. However, our results are in close agreement with Swanson's (1996) recent work, in which a different series of problems and individual difference measures of working memory was used. As appealing as the common resource notion is—and we too adopted this idea in our earlier work (Rabinowitz et al., 1989)—we must sound a note of caution. Specifically, common resource models often falter when developmental contrasts are conducted. That is, lifespan changes in reasoning– remembering tradeoffs in the solving of class-inclusion problems have resisted the more parsimonious single-resource interpretation (see Howe & Rabinowitz, 1996). Because models of cognitive performance, including resource models, can be critically evaluated in a developmental context (also see Karmiloff-Smith, 1992), it may be helpful to extend the Just and Carpenter (1992) simulation so that it incorporates dynamic growth parameters (see, e.g., Howe & Rabinowitz, 1994). This would be an important advance, because, as we have just seen, studies of a single age group, like the present one, may not provide data representative of the range of the cognitive process under scrutiny or the functioning of the system over the lifespan (Karmiloff-Smith, 1992). Such an extension may help us understand the complexities of resource management in problem-solving tasks such as these as well as the developmental pattern of reasoning– remembering relationships.

REFERENCES

Cantor, J., & Engle, R. W. (1993). Working-memory capacity as long-term memory activation: An individual-differences approach. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 1101-1114.

Conway, A. R. A., & Engle, R. W. (1994). Working memory and retrieval: A resource-dependent inhibition model. *Journal of Experimental Psychology: General*, **123**, 354-373.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*, **19**, 450-466.

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory & Language*, **25**, 1-18.

Hodkin, B. (1987). Performance analysis in class inclusion: An illustration with two language conditions. *Developmental Psychology*, **23**, 683-689.

Howe, M. L., & Rabinowitz, F. M. (1989). On the uninterpretability of dual-task performance. *Journal of Experimental Child Psychology*, **47**, 32-38.

Howe, M. L., & Rabinowitz, F. M. (1990). Resource panacea? Or just another day in the developmental forest. *Developmental Review*, **10**, 125-154.

Howe, M. L., & Rabinowitz, F. M. (1994). Dynamic modeling, chaos, and cognitive development. *Journal of Experimental Child Psychology*, **58**, 184-199.

Howe, M. L., & Rabinowitz, F. M. (1996). Reasoning from memory: A lifespan inquiry into the necessity of remembering when reasoning about class inclusion. *Journal of Experimental Child Psychology*, **61**, 1-42.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, **99**, 122-149.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research* (Vol. 4, pp. 1-18). New York: Plenum.

Perlmutter, N. J., & MacDonald, M. C. (1995). Individual differences and probabilistic constraints in syntactic ambiguity resolution. *Journal of Memory & Language*, **34**, 521-542.

Rabinowitz, F. M. (1995). Algorithm 744: A stochastic algorithm for global optimization with constraints. *ACM Transactions on Mathematical Software*, **21**, 194-213.

Rabinowitz, F. M., Howe, M. L., & Lawrence, J. A. (1989). Class inclusion and working memory. *Journal of Experimental Child Psychology*, **48**, 379-409.

Siddal, J. N., & Bonham, D. J. (1974). *Optimization subroutine package*. Hamilton, ON: McMaster University, Department of Mechanical Engineering.

Swanson, H. L. (1996). Individual and age-related differences in children's working memory. *Memory & Cognition*, **24**, 70-82.

Waters, G. S., & Caplan, D. (1996). Processing resource capacity and the comprehension of garden path sentences. *Memory & Cognition*, **24**, 342-355.

NOTES

1. As already mentioned, despite the fact that the problem information remained on the screen while subjects answered the questions, ceiling effects have not been obtained in the no load condition (Howe & Rabinowitz, 1996; Rabinowitz et al., 1989). As it turns out, ceiling effects were not present in this experiment either. We believe that the reason for this is that subjects are unable to read the problem information and reason at the same time. That is, all information must be in working memory for subjects to generate answers.

2. Technically, the 0 and 1 scores represent a qualitative difference in error type. Although this scoring scheme is arbitrary for the purposes of the ANOVA, it represents a meaningful distinction in the context of the mathematical model. A more traditional analysis in which all errors are scored as 0 and correct responses as 1 revealed a similar pattern of findings for the ANOVA.

3. Although we believe that the common resource is domain general, we cannot rule out the possibility that the resource is specific to language-based tasks.

APPENDIX A
The Materials Used in Each of the 48 Units

1. dogs, cats, animals, brown, and white
2. diamonds, squares, shapes, blue, and silver
3. stoves, toasters, appliances, beige, and green
4. colts, mares, horses, white, and brown
5. violins, guitars, instruments, black, and red
6. soup bowls, salad bowls, bowls, white, and blue
7. pliers, wrenches, tools, grey, and green

8. diamonds, rubies, jewels, white, and red
9. jets, gliders, planes, silver, and orange
10. gum drops, marshmallows, candy, pink, and white
11. peas, beans, vegetables, green, and yellow
12. pears, strawberries, fruit, green, and red
13. novels, dictionaries, books, blue, and black
14. jumbo jets, fighters, planes, pink, and silver
15. circles, triangles, shapes, pink, and purple
16. battleships, submarines, ships, grey, and black
17. peppermints, jelly beans, candy, pink, and green
18. silk pieces, linen pieces, cloth pieces, green, and blue
19. mixers, blenders, appliances, green, and pink
20. pianos, harps, instruments, black, and gold
21. pencils, pens, things to write with, yellow, and pink
22. cucumbers, pumpkins, vegetables, green, and orange
23. Buicks, Fords, cars, silver, and red
24. footballs, golf balls, balls, brown, and white
25. submarines, canoes, boats, yellow, and red
26. markers, crayons, things to write with, green, and purple
27. velvet pieces, denim pieces, cloth pieces, purple, and orange
28. ladybugs, wasps, insects, orange, and yellow
29. rings, bracelets, jewelry, gold, and silver
30. textbooks, comic books, books, green, and red
31. skyscrapers, castles, buildings, blue, and grey
32. tables, chairs, pieces of furniture, white, and brown
33. roses, lilies, flowers, red, and orange
34. robins, swallows, birds, red, and brown
35. Toyotas, Hondas, cars, grey, and blue
36. jays, crows, birds, blue, and black
37. teapots, coffee pots, pots, orange, and beige
38. marbles, blocks, toys, purple, and blue
39. sofas, chairs, pieces of furniture, red, and purple
40. saws, hammers, tools, brown, and silver
41. beetles, butterflies, insects, black, and yellow
42. pigs, cows, animals, pink, and brown
43. sapphires, emeralds, jewels, blue, and green
44. lilacs, primroses, flowers, white, and yellow
45. stallions, mares, horses, black, and brown
46. schools, jails, buildings, red, and black
47. lemons, apples, pieces of fruit, yellow, and green
48. watches, chains, jewelry, silver, and gold

APPENDIX B

The Choice Model Equations

Subclass–subclass problems. The probability of correctly answering a subclass–subclass problem, if the cues associated with each subclass were different, is equal to the probability of correctly encoding the cue values as different, multiplied by the conditional probability of correctly associating cue values and subclasses,

$$P(S_{ss}) = [ed]. \qquad \text{(B1, } nd, c\text{)}$$

An E_1 subclass–subclass error would accurately reflect labeling the cues as different, but reversing the association of cues and subclasses,

$$P(E_{1ss}) = [e(1 - d)]. \qquad \text{(B2, } nd, c\text{)}$$

An E_2 subclass–subclass error would reflect labelling the cues as same rather than different,

$$P(E_{2ss}) = [(1 - e)]. \qquad \text{(B3, } nd, c\text{)}$$

If the same cue was associated with each subclass, the probability of correctly answering a subclass–subclass problem is e,

$$P(S_{ss}) = [e]. \qquad \text{(B4, } ns\text{)}$$

In this case, E_1 and E_2 errors would be indistinguishable, and both would result from encoding the cues as different,

$$P(E_{1ss}) = P(E_{2ss}) = [(1 - e)/2]. \qquad \text{(B5, } ns\text{)}$$

Minor-subclass versus class problems. Minor-subclass versus class number–equal problems cannot be constructed. A minor–subclass versus class number-different or color problem can be answered correctly in a number of ways. If the subclasses are appropriately encoded as different and the cue values are remembered, then subclass–subclass and correct interpretations always result in correct solutions. The logic following either type of encoding error is different for number-different and color problems. If a subject understands class inclusion, then, as long as neither of the subclasses is empty, the relative numerosity of subclasses is irrelevant. The superordinate class is always larger than either subclass and the minor-subclass versus class number-based problem will always be solved correctly. On the other hand, if cue values are reversed or treated as the same with the minor-subclass versus class color-based problem, the subject who understands class inclusion will always respond "same color" and make an E_2 error. For example, if swallows are erroneously encoded as red rather than brown, the reddest bird is the "same color" as the reddest swallow.

For all problems involving class comparisons with a subclass, idiosyncratic interpretations generate correct solutions one third of the time, E_1 errors one third of the time, and E_2 errors one third of the time, independently of the way information has been encoded. Thus, "$i/3$" multiplies the probabilities associated with each of the possible encodings in all of the equations that follow:

$$P(S_{msc}) = [ed(u + s + i/3) + e(1 - d)(u + i/3) + (1 - e)(u + i/3)], \qquad \text{(B6, } nd\text{)}$$

$$P(S_{msc}) = [ed(u + s + i/3) + e(1 - d)i/3 + (1 - e)i/3]. \qquad \text{(B7, } c\text{)}$$

Note that in each of the equations that appear below, as well as in Equations B6 and B7, the terms are organized so that the term reflecting correct encoding (ed for the number-different and color problems, e for the number-same problems) appears first, whereas the term reflecting reversed same-different encoding ($1 - e$ for all problems) appears last.

In the minor-subclass versus class number-different and color problems, E_1 errors occur following subclass–subclass interpretations if the cue values associated with each subclass are reversed,

$$P(E_{1msc}) = [edi/3 + e(1 - d)(s + i/3) + (1 - e)i/3]. \qquad \text{(B8, } nd, c\text{)}$$

In the minor-subclass versus class number-different and color problems, E_2 errors (i.e., same-number or same-color choices) occur following "same" encoding and subclass–subclass interpretations. In addition, as explained earlier, E_2 errors will occur in color problems following any encoding error and correct class-inclusion interpretation,

$$P(E_{2msc}) = [edi/3 + e(1 - d)i/3 + (1 - e)(s + i/3)], \qquad \text{(B9, } nd\text{)}$$

$$P(E_{2msc}) = [edi/3 + e(1 - d)(u + i/3) + (1 - e)(u + s + i/3)]. \qquad \text{(B10, } c\text{)}$$

Class-inclusion problems. For all problem types, correct encoding of the cues followed by the correct interpretation of the class-inclusion problem results in a correct solution. Because the number of items is irrelevant, as long as neither subclass is empty, in number-based class-inclusion problems, understanding class inclusion always results in correct responding in number-different and number-same problems. In number-different problems, subclass–subclass interpretations of the class-inclusion problem result in correct answers if the cues are correctly encoded as different, and associated with the wrong subclasses. In the number-same problems, subclass–subclass interpretations of the class-inclusion problem result in correct solutions half the time if the cues erroneously are encoded as different, because it is assumed that each of the two cues is equally likely to be encoded as the more numerous. In the color problems, either correct or subclass–subclass interpretations of the class-inclusion problem result in correct solutions (i.e., "same color") following encoding of the cues as the same. Therefore,

$$P(S_{ci}) = [ed(u + i/3) + e(1 - d)(u + s + i/3) + (1 - e)(u + i/3)], \quad \text{(B11, nd)}$$

$$P(S_{ci}) = [ed(u + i/3) + e(1 - d)i/3 + (1 - e)(u + s + i/3)], \quad \text{(B12, c)}$$

$$P(S_{ci}) = [e(u + i/3) + (1 - e)(u + s/2 + i/3)]. \quad \text{(B13, ns)}$$

The equation for E_1 errors is identical for number-different and color problems with class inclusion. For all problem types, if the cues are correctly encoded, errors follow subclass–subclass interpretations,

$$P(E_{1ci}) = [ed(s + i/3) + e(1 - d)i/3 + (1 - e)i/3], \quad \text{(B14, nd, c)}$$

$$P(E_{1ci}) = [e(s + i/3) + (1 - e)i/3]. \quad \text{(B15, ns)}$$

For number-different problems with class inclusion, E_2 errors (i.e., "same number") result from encoding the cues as the same and subclass–subclass interpretations. For color problems, E_2 errors (i.e., "class y-er") follow correct encoding of the cues as different, reversing the cues associated with the two subclasses, and either correct or subclass–subclass interpretations of the class-inclusion problem. Note that the E_2 class-inclusion color error differs from all other types of E_2 errors in that it *cannot* follow same–different encoding errors (i.e., "same" color encoding, $1 - e$) unless the encoding error is associated with idiosyncratic interpretation ($i/3$). For number-same problems, E_2 errors (i.e., "subclass larger") occur half the time following erroneous encoding of the cues as different and subclass–subclass interpretations. Therefore,

$$P(E_{2ci}) = [edi/3 + e(1 - d)i/3 + (1 - e)(s + i/3)], \quad \text{(B16, nd)}$$

$$P(E_{2ci}) = [edi/3 + e(1 - d)(u + s + i/3) + (1 - e)i/3], \quad \text{(B17, c)}$$

$$P(E_{2ci}) = [ei/3 + (1 - e)(s/2 + i/3)]. \quad \text{(B18, ns)}$$

The Likelihood Functions

Different likelihood functions were used for the number-same, number-different, and color problems. This is because different data and different theoretical equations are relevant to each of these questions. The likelihood functions are stated according to the correct responses and errors associated with each question type. For each question type, there is an empirical likelihood associated with the data space, one that exhausts all of the empirical information, and a theoretical likelihood based on the equations developed in the preceding section of this appendix, one whose degrees of freedom (i.e., number of parameters estimated) are less than that found in the data space.

To begin, consider the empirical likelihood function for the number same questions,

$$\mathbf{L}_4 = P(S_{ss})^{N1} P(E_{1ss})^{N2} P(E_{2ss})^{N3} P(S_{ci})^{N4} P(E_{1ci})^{N5} P(E_{2ci})^{N6}, \quad \text{(B19, ns)}$$

where the Ps are observed proportions of either successes or error types and the Ns are the observed number of times each the events occurred. The theoretical likelihood function for the number-same problems is given by the same equation, except that the observed proportions are replaced by the equations describing the theoretical probabilities for each of the events. For example, e is substituted for $P(S_{ss})$; see Equation B4. Similarly $[ei/3 + (1 - e)(s/2 + i/3)]$ is substituted for $P(E_{2ci})$; see Equation B18. Because there are only three theoretical parameters involved, the theoretical likelihood function uses three degrees of freedom ($\mathbf{L}_3$).

The same process is used for number-different and color problems. The empirical likelihood function for the number different problems is given by

$$\mathbf{L}_8 = P(S_{ss})^{N1} P(E_{1ss})^{N2} P(E_{2ss})^{N3} P(S_{msc})^{N4+N4'} P(E_{1msc})^{N5+N5'} P(E_{2msc})^{N6+N6'} P(S_{ci})^{N7} P(E_{1ci})^{N8} P(E_{2ci})^{N9}, \quad \text{(B20, nd)}$$

where the Ps and Ns are as before and the primed N terms refer to the number-different problems substituted for the impossible number-same minor subclass–class problems. The theoretical likelihood function for the number-different problems uses four parameters ($\mathbf{L}_4$).

The empirical likelihood function for the color problems is given by

$$\mathbf{L}_6 = P(S_{ss})^{N1} P(E_{1ss})^{N2} P(E_{2ss})^{N3} P(S_{msc})^{N4'} P(E_{1msc})^{N5'} P(E_{2msc})^{N6'} P(S_{ci})^{N7} P(E_{1ci})^{N8} P(E_{2ci})^{N9}, \quad \text{(B21, c)}$$

where the theoretical likelihood function for the color problems also uses four parameters ($\mathbf{L}_4$).

(Manuscript received October 28, 1996;
revision accepted for publication July 10, 1997.)