Psychological Review
1990, Vol. 97, No. 3, 404–431

What One Intelligence Test Measures: A Theoretical Account of the Processing in the Raven Progressive Matrices Test

Patricia A. Carpenter, Marcel Adam Just, and Peter Shell
Carnegie Mellon University

The cognitive processes in a widely used, nonverbal test of analytic intelligence, the Raven Progressive Matrices Test (Raven, 1962), are analyzed in terms of which processes distinguish between higher scoring and lower scoring subjects and which processes are common to all subjects and all items on the test. The analysis is based on detailed performance characteristics, such as verbal protocols, eye-fixation patterns, and errors. The theory is expressed as a pair of computer simulation models that perform like the median or best college students in the sample. The processing characteristic common to all subjects is an incremental, reiterative strategy for encoding and inducing the regularities in each problem. The processes that distinguish among individuals are primarily the ability to induce abstract relations and the ability to dynamically manage a large set of problem-solving goals in working memory.

In this article, we analyze a form of thinking that is prototypical of what psychologists consider to be analytic intelligence. We use the term *analytic intelligence* to refer to the ability to reason and solve problems involving new information, without relying extensively on an explicit base of declarative knowledge derived from either schooling or previous experience. In the theory of R. Cattell (1963), this form of intelligence has been labeled *fluid intelligence* and has been contrasted with *crystallized intelligence*, which more directly reflects the previously acquired knowledge and skills that have been crystallized with experience. Thus, analytic intelligence refers to the ability to deal with novelty, to adapt one's thinking to a new cognitive problem. In this article, we provide a theoretical account of what it means to perform well on a classic test of analytic intelligence, the Raven Progressive Matrices Test (Raven, 1962).

We describe a detailed theoretical model of the processes used in solving the Raven test, contrasting the performance of college students who are less successful in solving the problems with those who are more successful. The model is based on multiple dependent measures, including verbal reports, eye fixations, and patterns of errors on different types of problems. The experimental investigations led to the development of computer simulation models that test the sufficiency of our analysis. Two computer simulations, FAIRAVEN and BETTERAVEN, express the differences between good and extremely good performance on the test. The FAIRAVEN model performs like the median college student in our sample; BETTERAVEN performs like one of the very best. The BETTERAVEN model differs from FAIRAVEN in two major ways: BETTERAVEN has the ability to induce more abstract relations than FAIRAVEN, and BETTERAVEN has the ability to manage a larger set of goals in working memory and hence can solve more complex problems. The two models and the contrast between them specify the nature of the analytic intelligence required to perform the test and the nature of individual differences in this type of intelligence.

There are several reasons why the Raven test provides an appropriate test bed to study analytic intelligence. First, the size and stability of the individual differences that the test elicits, even among college students, suggest that the underlying differences in cognitive processes are susceptible to cognitive analysis. Second, the relatively large number of items on the test (36 problems) permits an adequate data base for the theoretical and experimental analyses of the problem-solving behavior. Third, the visual format of the problems makes it possible to exploit the fine-grained, process-tracing methodology afforded by eye-fixation studies (Just & Carpenter, 1976). Finally, the correlation between Raven test scores and measures of intellectual achievement suggests that the underlying processes may be general rather than specific to this one test (Court & Raven, 1982), although like most correlations, this one must be interpreted with caution.

The Raven test, including the simpler Standard Progressive Matrices Test and the Coloured Progressive Matrices Test, is also widely used in both research and clinical settings. The test is used extensively by the military in several Western countries (for example, see Belmont & Marolla, 1973). Also, because of its nonverbal format, the test is a common research tool used with children, the elderly, and patient populations for whom the processing of language may need to be minimized. The wide usage means that there is a great deal of information about the

This work was supported in part by Contract N00014-89-J-1218 from the Office of Naval Research and by Research Scientist Development Awards MH-00661 and MH-00662 from the National Institute of Mental Health. The simulation models were programmed by Peter Shell and Craig Bearer.

We are grateful to Earl Hunt, David Klahr, Kenneth Kotovsky, Allen Newell, James Pellegrino, Herbert Simon, and Robert Sternberg for their constructive comments on drafts of the manuscript. We also want to acknowledge the help of David Fallside in implementing and analyzing the Tower of Hanoi experiment.

Correspondence concerning this article should be addressed to Patricia A. Carpenter, Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

performance profiles of various populations. But more important, it means that a cognitive analysis of the processes and structures that underlie performance has potential practical importance in the domains in which the test is used for either research or classification.

Several different research approaches have converged on the conclusion that the Raven test measures processes that are central to analytic intelligence. Individual differences in the Raven test correlate highly with those found in other complex, cognitive tests (see Jensen, 1987). The centrality of the Raven test among psychometric tests is graphically illustrated in several nonmetric scaling studies that examined the interrelations among ability test scores obtained from archival sources and from more recently collected data (Snow, Kyllonen, & Marshalek, 1984). The scaling solutions for the different data bases showed remarkably similar patterns. The Raven test and other complex reasoning tests were at the center of the solution. Simpler tests were located toward the periphery, and they clustered according to their content, as shown in Figure 1 (top panel). This particular scaling analysis is based on the results from various cognitive tests given to 241 high school students (Marshalek, Lohman, & Snow, 1983). Snow et al. constructed an idealized space to summarize the results of their numerous scaling solutions, in which they placed the Raven test at the center, as shown in Figure 1 (bottom panel). In this idealized solution, task complexity is maximal near the center and decreases outward, toward the periphery. The tests in the annulus surrounding the Raven test involve abstract reasoning, induction of relations, and deduction. For tests of intermediate or low complexity only, there is clustering as a function of the test content, with separate clusters for verbal, numerical, and spatial tests. By contrast, the more complex tests of reasoning at the center of the space were highly intercorrelated in spite of differences in specific content.

One of the sources of the Raven test's centrality, according to Marshalek et al. (1983), is that "more complex tasks may require more involvement of executive assembly and control processes that structure and analyze the problem, assemble a strategy of attack on it, monitor the performance process, and adapt these strategies as performance proceeds" (p. 124). This theoretical interpretation is based on the outcome of the scaling studies. Our research also converges on the importance of executive processes, but the conclusions are derived from a process analysis of the Raven test.

Although there has been some dispute among psychometricians about which tests in the larger space might be said to reflect analytic intelligence, the Raven test is central with respect to either interpretation. In one view, intelligence refers to a construct underlying a small range of tests, namely those at the center of the space. This view is associated with Spearman (1927), although Spearman himself avoided the term *intelligence* and instead used the term *g* to refer to the determinants of shared variance among tests of intellectual ability (Jensen, 1987). An alternative view, associated with Thurstone (1938), applies the term *intelligence* to a large set of diverse mental abilities, including not only those at the center of the space but also some domain-specific abilities, such as those in the periphery of the space. Although the two views differ in the size of the spaces which they associate with intelligence, the centrality of the Raven test emerges in either case. The centrality of the Raven test indicates not only that it is a good measure of intelligence, but also that a theory of the processing in the Raven test should account for a good deal of the reasoning in the other tests in the center of the space.

This article is organized in four parts. The structure of the problems is described in the Problem Structure and Human Performance section, which focuses on the problem characteristics that are likely to tax the psychological processes. We also report two studies that examine the processes empirically, determining which processes distinguish between high-scoring subjects and lower scoring subjects and which processes are common to all subjects in their attempts to solve all problems. In the Simulation Models section, we describe the two simulation models that perform like the median subject or like the best subject. Next, we compare the performance of the human subjects and the theoretical models in detail in Comparing Human Performance to the Theory. In the final section, Cognitive Processes and Human Intelligence, we generalize the theory and examine its implications for a theory of intelligence.

Problem Structure and Human Performance

A task analysis of the Raven Progressive Matrices Test suggests some of the cognitive processes that are likely to be implicated in solving the problems. The test consists of a set of visual analogy problems. Each problem consists of a 3 × 3 matrix, in which the bottom right entry is missing and must be selected from among eight response alternatives arranged below the matrix. (Note that the word *entry* refers to each of the nine cells of the matrix.) Each entry typically contains one to five figural elements, such as geometric figures, lines, or background textures. The test instructions tell the test taker to look across the rows and then look down the columns to determine the rules and then to use the rules to determine the missing entry. The problem in Figure 2 illustrates the format.[1]

The variation among the entries in a row and column of this problem can be described by three rules:

> *Rule A.* Each row contains three geometric figures (a diamond, a triangle, and a square) distributed across its three entries.
>
> *Rule B.* Each row contains three textured lines (dark, striped, and clear) distributed across its three entries.
>
> *Rule C.* The orientation of the lines is constant within a row but varies between rows (vertical, horizontal, and oblique).

The missing entry can be generated from these rules. Rule A specifies that the answer should contain a square (because the first two columns of the third row contain a triangle and diamond). Rule B specifies it should contain a dark line. Rule C specifies that the line orientation should be oblique, from upper left to lower right. These rules converge on the correct response alternative, 5. Some of the incorrect response alternatives are

[1] To protect the security of the Raven problems, none of the actual problems from the test are depicted here or elsewhere in this article. Instead, the test problems are illustrated with isomorphs that use the same rules but different figural elements and attributes. The actual problems that were presented to the subjects are referred to by their number in the test, which can be consulted by readers.

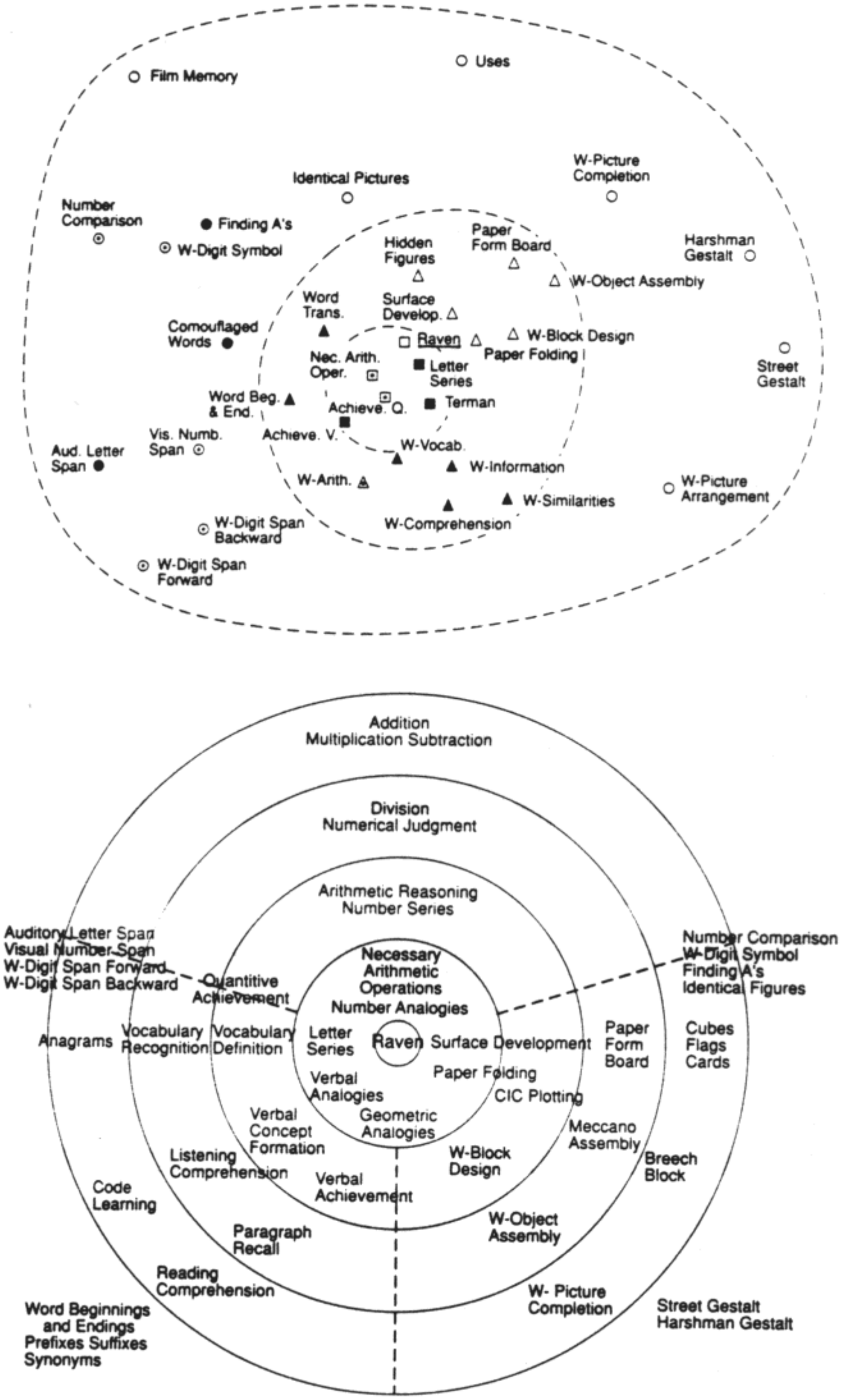

Figure 1. Scalings of the intercorrelations among various ability tests showing the centrality of the Raven test. (Tests near the center of the space, such as the Raven test and letter-series test, are the most complex and share considerable variance in spite of their differences in content [figural vs. verbal]. The outwardly radiating concentric circles indicate decreasing levels of complexity and show increasing separation as a function of test content. Top panel: Nonmetric scaling of the data from 241 high school students. Test complexity is indicated by the shapes of the points [squares, most complex; triangles, intermediately complex; circles, least complex]. W = Wechsler Adult Intelligence Scale. Bottom panel: An idealization of the analyses of several psychometric batteries. Tests involving different content—figural, verbal, and numerical—are separated by dashed lines. *Note:* The top panel is from "The Complexity Continuum in the Radex

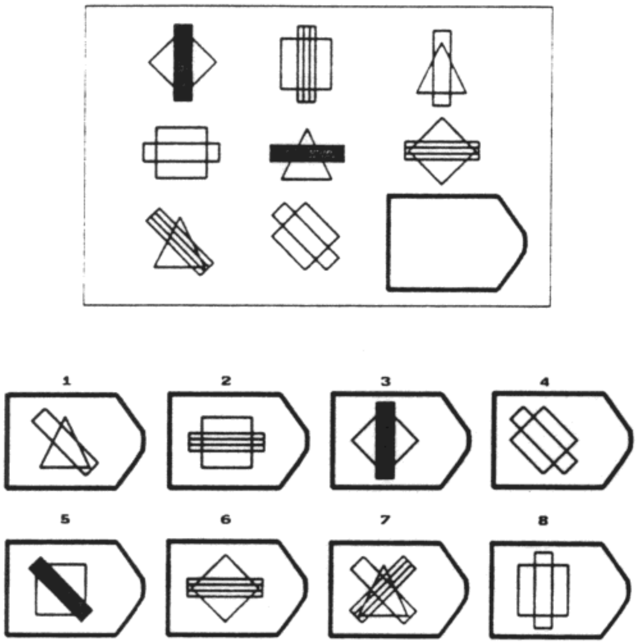

Figure 2. A problem to illustrate the format of the Raven items. (The variation among the three geometric forms [diamond, square, and triangle] and three textures of the line [dark, striped, and clear] is each governed by a distribution-of-three-values rule. The orientation of the line is governed by a constant-in-a-row rule. The correct answer is 5.)

designed to satisfy an incomplete set of rules. For example, if a subject induced Rule A but not B or C, he or she might choose Alternative 2 or 8. Similarly, inducing Rule B but omitting A and C leads to Alternative 3. This sample problem illustrates the general structure of the test problems but corresponds to one of the easiest problems in the test. The more difficult problems entail more rules or more difficult rules, and more figural elements per entry.

Our research focuses on a form of the Raven test that is used widely for adults of higher ability, the Raven Advanced Progressive Matrices, Sets I and II. Set I, consisting of 12 problems, is often used as a practice test or to obtain a rough estimate of a subject's ability. The first several problems in Set I can be solved by perceptually based algorithms such as line continuation (Hunt, 1974). However, the later problems in Set I and most of the 36 problems in Set II, which our research examines, cannot be solved by perceptually based algorithms, as Hunt noted. Like the sample problem in Figure 2, the more difficult problems require that subjects analyze the variation in the problem to induce the rules that generate the correct solution. The problems requiring an analytic strategy can be used to discriminate among individuals with higher education, such as college students (Raven, 1965).

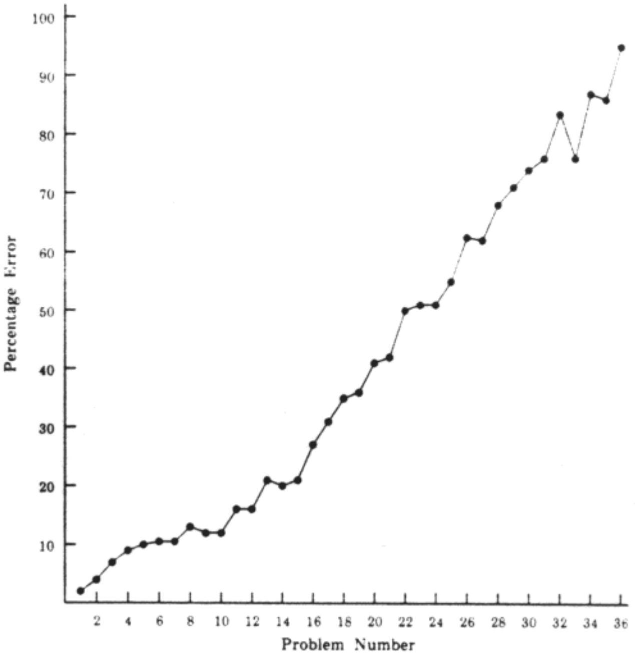

Figure 3. The percentage error for each problem in Set II of the Raven Advanced Progressive Matrices Test shows the large variation in difficulty among problems with very similar formats. (The data are from 2,256 British adults, including telephone engineering applicants, students at a teacher training college, and British Royal Air Force recruits [Forbes, 1964].)

Problem Difficulty

Although all of the Raven problems share a similar format, there is substantial variation among them in their difficulty. The magnitude of the variation is apparent from the error rates (shown in Figure 3) of 2,256 British adults, including telephone engineering applicants, students at a teacher training college, and British Royal Air Force recruits (Forbes, 1964). There is an almost monotonic increase in difficulty from the initial problems, which have negligible error rates, to the last few problems, which have extremely high error rates. (The error rates on the final problems reflect failures to attempt these problems in the testing period as well as failures to solve them correctly.) The considerable range of error rates among problems leads to the question of what psychological processes account for the differences in problem difficulty and for the differences among people in their ability to solve them.

The test's origins provide a clue to what the test was intended to measure. The Raven Progressive Matrices Test was developed by John Raven, a student of Spearman. As we previously mentioned, Spearman (1927) believed that there was one central intellectual ability (which he referred to as *g*), as well as numer-

and Hierarchical Models of Intelligence" by B. Marshalek, D. F. Lohman, and R. E. Snow, 1983, *Intelligence, 7,* p. 122. Copyright 1983 by Erlbaum. Reprinted by permission. The bottom panel is from *Advances in the Psychology of Human Intelligence* [p. 92] by R. E. Snow, P. C. Kyllonen, and B. Marshalek, 1984. Hillsdale, NJ:Erlbaum. Copyright 1984 by Erlbaum. Reprinted by permission.)

ous specific abilities. He never precisely defined what *g* consisted of, but it was thought to involve "the eduction of relations and correlates" (Spearman, 1927, pp. 165–166). Raven's conception of what his progressive matrices test measured was somewhat more articulated. His personal notes, generously made available to us by his son, J. Raven, indicate that he wanted to develop a series of overlapping, homogeneous problems whose solutions required different abilities. However, the descriptions of the abilities that Raven intended to measure are primarily characteristics of the problems, not specifications of the requisite cognitive processes. John Raven constructed problems that focused on each of six different problem characteristics, which approximately correspond to the different types of rules that we describe later. He used his intuition and clinical experience to rank order the difficulty of the six problem types. Many years later, normative data from Forbes (1964), shown in Figure 3, became the basis for selecting problems for retention in newer versions of the test and for arranging the problems in order of increasing difficulty, without regard to any underlying processing theory. Thus, the version of the test that is examined in this research is an amalgam of Raven's implicit theory of the components of reasoning ability and subsequent item selection and ordering done on an actuarial basis.

Table 1
A Taxonomy of Rules in the Raven Test

Rule	Taxonomy
Constant in a row	The same value occurs throughout a row, but changes down a column. (See Figure 4b, where the location of the dark component is constant within each row; in the top row, the location is the upper half of the diamond; in the middle row, it is the bottom half of the diamond; and in the bottom row, it is both halves.)
Quantitative pairwise progression	A quantitative increment or decrement occurs between adjacent entries in an attribute such as size, position, or number. (See Figure 4a, where the number of black squares in each entry increases along a row from 1 to 2 to 3.)
Figure addition or subtraction	A figure from one column is added to (juxtaposed or superimposed) or subtracted from another figure to produce the third. (See Figure 4b, where the figural element in column 1 juxtaposed to the element in column 2 produces the element in column 3.)
Distribution of three values	Three values from a categorical attribute (such as figure type) are distributed through a row. (See Figure 2, where the three geometric forms—diamond, square, and triangle—follow a distribution rule and the three line textures—black, striped, and clear—also follow a distribution rule.)
Distribution of two values	Two values from a categorical attribute are distributed through a row; the third value is null. (See Figure 4c, where the various figural elements, such as the vertical line, the horizontal line, and the V in the first row, follow a distribution of two values.)

Rule Taxonomy

Across the Raven problems that we have examined, we found that five different *types* of rules govern the variation among the entries. Many problems involve multiple rules, which may all be different rule types or several instances or *tokens* of the same type of rule. Table 1 shows the five types of rules that are illustrated by the problems in Figures 2 and 4. Almost all of the Raven problems in Sets I and II can be classified with respect to which of these rule types govern its variation, as shown in the Appendix.[2]

One qualification to this analysis is that sometimes the set of rules describing the variation in a problem is not unique. For example, quantitative pairwise progression is often interchangeable with a distribution-of-three-values rule. Consider a row consisting of three arrows pointing to 12, 4, and 8 o'clock. This variation can be described as a distribution of three values or in terms of a quantitative progression, in which the arrow's orientation is progressively rotated 120° clockwise, beginning at 12 o'clock. Similarly, the variation described by a distribution-of-two-values rule may be alternatively described by a figure-addition-modulo-2 rule. In the case of alternative rules, the Appendix lists the rules most often mentioned by the highest scoring subjects in Experiment 1a.[3]

Finding Corresponding Elements

In problems with multiple rules, the problem solver must determine which figural elements or attributes in the three entries in a row are governed by the same rule, a process that will be called *correspondence finding.* For example, given a shaded square in one entry, the problem solver might have to decide which figure in another entry, either a shaded triangle or an unshaded square, is governed by the same rule. Do the squares correspond to each other, or do the shaded figures? In this exam-

[2] This analysis is row oriented. In most problems, the rule types are the same regardless of whether a row or column organization is applied; in our experiments, we found that most subjects analyzed the problems by rows. Two of the problems on the test were unclassifiable within our taxonomy because the nature of their rules differed from all others.

[3] This taxonomy finds some converging support from an analysis of the relations used in figural analogies (both 2 × 2 and 3 × 3 matrices) from 166 intelligence tests (Jacobs & Vandeventer, 1972). Jacobs and Vandeventer found that 12 relations accounted for many of the analogical problems. Five of their relations are closely related to rules we found in the Raven test: addition and added element (addition or subtraction), elements of a set (distribution of three values), unique addition (distribution of two values), and identity (constant in a row). Some of the remaining relations, such as numerical series and movement in a plane, map onto our quantitative progression rule. The Jacobs and Vandeventer analysis suggests that relatively few relations are needed to describe the visual analogies in a large number of such tests.

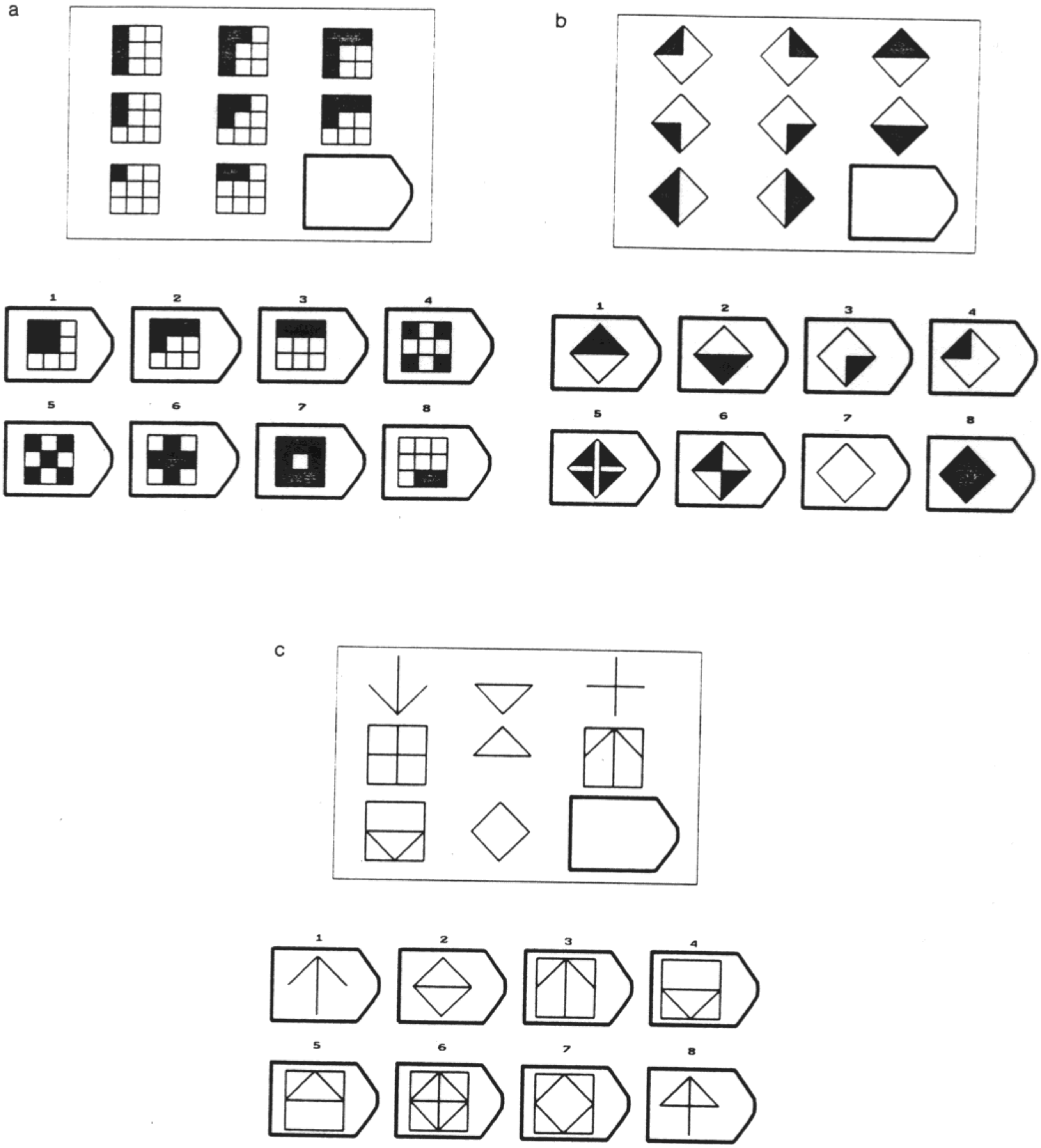

Figure 4. Problems illustrating rules of the Raven test. (Panel a: The quantitative pairwise progression rule. The number of black squares in the top of each row increases by one from the first to the second column and from the second to the third column. The number of black squares along the left remains constant within a row but changes between rows from three to two to one. The correct answer is 3. Panel b: The figure addition rule. The figural element in the first column is superimposed on the figural element in the second column to compose the figural element in the third column. The position of the darkened element remains constant in a row but changes between rows from top to bottom to both. The correct answer is 8. Panel c: The distribution-of-two-values rule. Each figural element, such as the horizontal line, the vertical line, the V, and so on, occurs twice in a row, and the third value is null. The correct answer is 5.)

ple, and in some of the Raven problems, the cues to the correspondence are ambiguous, making it difficult to tell a priori which figural elements correspond to each other. The correspondence-finding process is a subtle source of difficulty because many problems seem to have been constructed by conjoining the figural elements governed by several rules, without much regard for the possible difficulty of conceptually segmenting the conjunction.

The difficulty in correspondence finding can be illustrated with an adaptation of one of the problems (Set II, Problem 28), shown in Figure 5. A first plausible hypothesis about the correspondences is that the rectangles are governed by one rule, the curves by another rule, and the straight lines by a third rule. This hypothesis reflects the use of a *matching-names heuristic,* namely that figures with the same name might correspond to each other. If this hypothesis is pursued further, it becomes clear that although each row contains two instances of each figure type, the number and orientation of the figures vary unsystematically. The matching-names heuristic produces an unfruitful hypothesis about the correspondences in this problem. A subject who has tried to apply the heuristic must backtrack and consider other correspondences that are based on some other feature, either number or orientation. Number, like figure identity, does not result in any economical and complete rule that governs location or orientation. Orientation, the remaining attribute, is the basis for two economical, complete rules. The horizontal elements in each row can be described in terms of two distribution-of-three-values rules, one governing number (one, two, and three elements) and the other governing figure type (line, curve, and rectangle). Similarly, the vertical elements in each row are governed by the same two rules. This example illustrates the complexity of correspondence finding, which, along with the type of rule in a problem and the number of rules, can contribute to the difficulty of a problem.

In addition to variation among problems in the difficulty of correspondence finding, the problems also vary in the number of rules. Although Raven intended to evaluate a test taker's ability to induce relations, he apparently tried to make the induction process more difficult in some problems by including more examples or tokens of rules. A major claim of our analysis is that the presence of a larger number of rule tokens taxes not so much the processes that induce the rules, but the goal-management processes that are required to construct, execute, and maintain a mental plan of action during the solution of those problems containing multiple rule tokens as well as difficult correspondence finding.

Experiment 1: Performance in the Raven Test

The purpose of Experiment 1 was to collect more detailed data about performance in the Raven test to reveal more about the process and the content of thought during the solving of each Raven problem. Three types of measures, obtained in Experiments 1a and 1b, provided the basis for the quantitative evaluation of the theory.

The first measure was the frequency and pattern of errors, which were obtained in Experiments 1a and 1b. The simulation models account not only for the number of errors that a person of a given ability will make but also predict which types of problems he or she will fail to solve. The second type of measure, obtained in Experiment 1a, reflects on-line processes used during problem solution. One such on-line measure assessed how the entries in successive rows were visually examined. In particular, measures of the eye-fixation patterns assessed the number of times a subject scanned a row of entries and the number of times he or she looked back and forth (made paired comparisons) between entries. Another on-line measure was the time between the successive statements of rules uttered by subjects who were talking aloud while solving the problems. These online measures constrain the type of solution processes postulated in the simulations. The third measure, obtained in Experiment 1b, was the subjects' descriptions of the rules that they induced in choosing a response to each problem. The subjects' rules were compared with the rules induced by the simulation models.

Method

Procedure for Experiment 1a. In Experiment 1a, the subjects were presented with problems from the Raven test while their eye fixations were recorded. They were asked to talk out loud while they solved the problems, describing what they noticed and what hypotheses they were entertaining. The subjects were given the standard psychometric in-

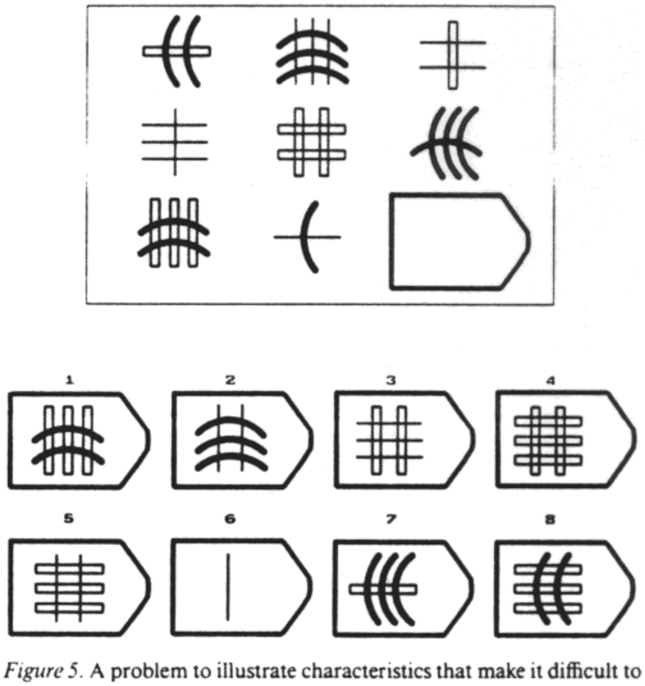

Figure 5. A problem to illustrate characteristics that make it difficult to determine which figural elements correspond, that is, which are operated on by the same rule. (Subjects initially assume that the rectangles correspond to each other, the dark curves correspond to each other, and the straight lines correspond to each other. But to solve the problem, subjects must backtrack and try other possible bases for correspondence, such as numerosity or orientation. Orientation turns out to provide the correct basis. The horizontal figures correspond to each other; their form [rectangle, curve, and straight line] and number [1, 3, and 2] are governed by distribution-of-three-values rules. Similarly, the vertical figures correspond to each other; their form and number are also governed by distribution-of-three-values rules. The correct answer is 5.)

structions and shown two simple practice problems. One deviation from standard psychometric procedure was that subjects were told to pace themselves so as to attempt all of the problems in the standard 40-min time limit.

Stimuli. Experiment 1a used 34 of the 48 problems in Sets I and II that could be represented and displayed within the raster graphics resolution of our display system, which was 512 × 512 pixels (see Just & Carpenter, 1979, for a description of the video digitization and display characteristics). The stimuli were created by digitizing the video image of each problem in the Raven test booklet. The Appendix shows the sequence number in the Raven test of the problems that were retained. The problems that could not be adequately digitized were those with very high spatial frequencies in their depiction, such as small grids or crosshatching (Set II, Problems 2, 11, 15, 20, 21, 24, 25, 28, and 30). There was little relation between the presence of high spatial frequencies and a problem's difficulty, as indicated by the normative error rate from Forbes (1964) shown in Figure 3.

Eye fixations. The subjects' eye fixations were monitored remotely with an Applied Science Laboratories corneal and pupil-centered eye tracker that sampled at 60 Hz, ultimately resulting in an x–y pair of gaze coordinates expressed in the coordinate system of the display. The individual x–y coordinates were later aggregated into fixations. Then, successive fixations on the same one of the nine entries in the problem matrix or on a single response alternative were aggregated together into units called *gazes*, which constituted the main eye-fixation data base.

Procedure for Experiment 1b. Unlike Experiment 1a, in which subjects gave verbal protocols while they solved each problem, in Experiment 1b, subjects were asked to work silently, make their response, and then describe the rules that motivated their final response. This change in procedure was intended to provide more complete information about what rules the subjects induced. These rule statements were then compared with the rules induced by FAIRAVEN and BETTERAVEN. Subjects were given 40 problems, approximately half of which were from the Raven Progressive Matrices Test and half from the Standard Progressive Matrices Test, involving similar rule types, to increase the number of problems involving more difficult rules. The subjects in Experiment 1b were tested in two sessions separated by approximately 1 week, with 20 items in each session.

Subjects. In Experiment 1a, the subjects were 12 Carnegie Mellon students, who participated for course credit. In Experiment 1b, the subjects were 22 students from Carnegie Mellon and the University of Pittsburgh, who participated for a $10 payment. Data were not included from 3 additional subjects who did not return for the second session to complete Experiment 1b.

Overview of Results

This overview presents the general patterns of results, particularly results that influenced the design features of the simulation models. This overview, presented in preliminary and qualitative terms, is followed by a more precise analysis of the data (see Comparing Human Performance to the Theory) after the presentation of the models.

In Experiment 1a, over all 34 problems, the number of errors per subject ranged from 2 to 20, with a mean of 10.5 (31%) and a median of 10.3. Although our college student subjects had a lower mean error rate than Forbes's (1964) more heterogeneous sample, the correlation between the error rates of our sample and Forbes's on the 27 problems in Set II was high, $r(25) = .91$. In Experiment 1b, the mean number of errors for the 40 Raven problems was 11.1 (28%), with a median of 10 errors.[4]

The error rate on a given problem was related to the types of rules it involved and the number of tokens of each rule type. A simple linear regression with a single independent variable that coded the total number of rules in a problem (regardless of whether they were of similar or different types and excluding any constant rules) accounted for 57% of the variance among the mean error rates in Experiment 1a for the 32 problems classified within our taxonomy. (If constant rules are included in the number of rule tokens in a problem, then the percentage of variance accounted for declines to 45%.) The median and mean response times for correct responses were generally longer for the problems that had higher error rates (with a correlation of .87 between the mean times and the errors), suggesting that problem difficulty affected both performance measures.

Perhaps the most striking facet of the eye fixations and verbal protocols was the demonstrably incremental nature of the processing. The way that the subjects solved a problem was to decompose it into successively smaller subproblems and then proceed to solve each subproblem. The induction of the rules was incremental in two respects. First of all, in problems containing more than one rule, the rules were described one at a time, with long intervals between rule descriptions, suggesting that they were induced one at a time. Second, the induction of each rule consisted of many small steps, reflected in the pairwise comparison of elements in adjoining entries. These aspects of incremental processing were ubiquitous characteristics of the problem solving of all of the subjects and do not appear to be a source of individual differences. Consequently, the incremental processing played a large role in the design of both simulation models.

A typical protocol (in Figure 6) from 1 subject illustrates the incremental processing. Table 2 shows the sequence of gazes and verbal comments made by an average subject (41% errors) solving a problem involving two distribution-of-three-values rules and a constant-in-a-row rule (Set II, Problem 1, which is isomorphic to the problem depicted in Figure 2). The subject's comments are transcribed adjacent to the gazes that occurred during the utterance. (The subject's actual comments were translated to refer to the isomorphic attributes depicted in Figure 2.) The location of each gaze is indicated by labeling the rows in the matrix from top to bottom as rows 1, 2, and 3 and the columns from left to right as 1, 2, and 3, such that (1, 2) designates the entry in the top row, middle column. The braces encompassing a sequence of gazes indicate how the gazes were classified in the analysis that counted the number of scans of rows and columns. The duration of each gaze is indicated in milliseconds next to the location of the gaze.

[4] A control study showed that the deviations from conventional administration did not alter the basic processing in Experiment 1a. A separate group of 19 college students was given the test without recording eye fixations or requiring concurrent verbal protocols. This control group produced the same pattern of errors, $r(25) = .93$, and response times, $r(25) = .89$, as the subjects in Experiment 1a for the 27 problems from Set II that both groups were presented. Furthermore, the error rate was slightly higher in Experiment 1a (33%) than in the control group (25%), demonstrating that the lower rate in Experiment 1a (and 1b) compared with Forbes's (1964) sample is probably due to our sample's comprising exclusively college students, rather than to increased accuracy in Experiment 1a because of eye-fixation recording or generation of verbal protocols.

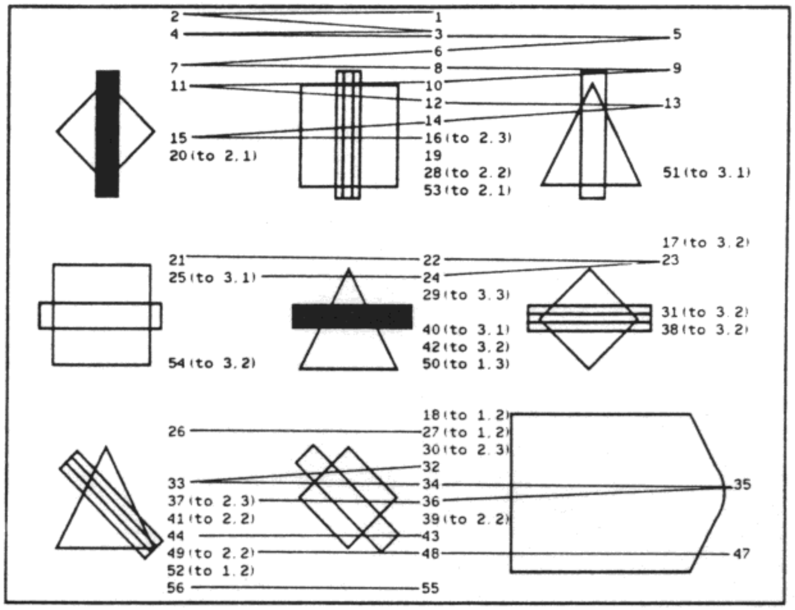

Figure 6. The sequence of gazes shown in the protocol in Table 2. (Gazes within the same row are connected. The numbers in parentheses indicate the location of a gaze that followed if it was in a different row.)

Table 2
Gazes in a Typical Protocol (Subject 5, Raven Set II, Problem 1)

Gaze no.	Location (row, column)		Duration (ms)	Subject's comments
1	Pairwise (1, 1)–(1, 2)	1, 2	233	
2		1, 1	367	
3		1, 2	533	
4	Row 1	1, 1	117	
5		1, 3	434	
6		1, 2	367	"Okay,
7	Row 1	1, 1	516	
8		1, 2	400	
9		1, 3	517	
10		1, 2	550	
11	Row 1	1, 1	383	there's diamond, square, triangle,
12		1, 2	517	
13		1, 3	285	
14	Pairwise (1, 1)–(1, 2)	1, 2	599	
15		1, 1	533	
16		1, 2	468	
17		2, 3	284	
18		3, 2	317	
19		1, 2	434	and they each contain lines through them
20		1, 1	533	
21	Row 2	2, 1	434	
22		2, 2	451	
23		2, 3	467	
24		2, 2	467	
25		2, 1	167	
26	Row 3	3, 1	233	
27		3, 2	267	with different shadings
28	Column 2	1, 2	483	
29		2.2	599	
30		3, 2	300	
31		2, 3	167	going from vertical, horizontal, oblique,
32	Row 3	3, 2	133	
33		3, 1	650	
34		3, 2	433	
35	Row 3	3, 3	432	
36		3, 2	167	
37		3, 1	400	
38		2, 3	217	and the third one should be—
39		3, 2	583	
40	Pairwise (2, 2)–(3, 1)	2, 2	583	
41		3, 1	334	
42		2, 2	267	
43	Row 3	3, 2	383	
44		3, 1	234	
45	Answers	7	183	
46		4	467	
47	Row 3	3, 3	217	
48		3, 2	199	
49		3, 1	183	
50	Diagonal	2, 2	350	
51		1, 3	150	
52		3, 1	433	
53		1, 2	234	
54		2, 1	117	Okay, it should be a square
55	Row 3	3, 2	417	
56		3, 1	366	
57	Answer	1	250	
58		5	1.900	and should have the
59		1	250	
60		5	184	black line in them and the answer's 5."

The verbal report shows that the subject mentioned one attribute at a time, with some time interval between the mentions, suggesting that the representation of the entries was being constructed incrementally. Also, the subject described the rules one at a time, typically with several seconds elapsing between rules. The subject seemed to construct a complete representation attribute by attribute and induced the rules one at a time.

The incremental nature of the process is also apparent in the pattern of gazes, particularly the multiple scans of rows and columns and the repeated fixations of pairs of related entries. These scans are apparent in the sequence of gazes shown in Figure 6. (The numbers indicating the sequence of gazes have been placed in columns to the right of the fixated entries, and lines have been drawn to connect the successive fixations of entries within rows.) This protocol indicates the large amount of pairwise and row-wise scanning. For example, like most of the eye-fixation protocols, this one began with a sequence of pairwise gazes on the first two entries in the top row. The subject was presumably encoding some of the figural elements in these two entries and comparing their attributes. The subject went on to compare middle and right entries of the top row, followed by several scans of the complete row.

The general results, then, are that the processing is incremental, that the number of rule tokens affects the error rates, and that there is a wide range of differences among individuals in their performance on this test.

Experiment 2: Goal Management in Other Tasks

The finding that error rates increase with the number of rule tokens in a problem suggests that the sheer keeping track of figural attributes and rules might be a substantial source of individual differences in the Raven test. "Keeping track" refers to the ability to generate subgoals in working memory, record the attainment of subgoals, and set new subgoals as others are attained. Subjects who are successful at goal management in the Raven test should also perform well on other cognitive tasks involving extensive goal management. One such task is a puzzle called the Tower of Hanoi, which can be solved with a strategy that requires considerable goal management. Most research on the Tower of Hanoi puzzle has focused on how subjects induce a correct strategy. By contrast, in the current study, the inductive aspect of the puzzle was minimized by teaching subjects a strategy beforehand, with extensive instructions and practice. Errors on the Tower of Hanoi puzzle should correlate with errors on the Raven test to the extent that both require goal management.

The Tower of Hanoi puzzle consists of three pegs and three or more disks of increasing size arranged on one of the pegs in the form of a pyramid, with the largest disk on the bottom and smallest disk on the top, as shown in the top corner of Figure 7. The subject's task is to reconstruct the pyramid, moving one disk at a time, onto another peg (the goal peg) without putting a larger disk on a smaller disk. One of the most commonly used strategies in the puzzle is called the *goal-recursion strategy* (Kotovsky, Hayes, & Simon, 1985; Simon, 1975). With this strategy, the puzzle is solved by first setting the goal of moving the largest disk from the bottom of the pyramid on the source peg to the goal peg. But before executing that move, the disks constituting the subpyramid above the largest disk must be moved out of the way. This goal is recursive because in order to move the subpyramid, its largest disk must be cleared, and so on. Thus, to execute this strategy, a subject must set up several embedded subgoals. As the number of disks in a puzzle increases, the subject must generate a successively larger hierarchy of subgoals and remember his or her place in the hierarchy while executing successive moves.

Moves in the Tower of Hanoi puzzle can be organized within a goal tree that specifies the subgoals that must be generated or retained on each move, as pointed out by Egan and Greeno (1974). Figure 7 shows a diagram of the goal tree when the goal-recursion strategy is used in a four-disk problem. Each branch corresponds to a subgoal, and the terminal nodes correspond to individual moves. The subject can be viewed as doing a depth-first search of the goal tree; in the goal-recursion strategy, the subject is taught to generate the subgoals equivalent to those listed in the leftmost branch to enable the first move. Subsequent moves entail maintaining, generating, or attaining various subgoals. In particular, on Moves 1, 5, 9, and 13, the claim is that the subject should generate one or more subgoals before executing the move; by contrast, no new subgoals need to be generated before other moves. Egan and Greeno found that likelihood of an error on a move increased with the number of goals that had to be maintained or generated to enable that move. Consequently, performance on the Tower of Hanoi goal-recursion strategy should correlate with performance on the Raven test, to the extent that both tasks rely on generating and maintaining goals in working memory.

Method

Procedure. The subjects were administered the Raven Progressive Matrices Test, Sets I and II, using standard psychometric procedures. Then the subjects were given extensive instruction and practice on the goal-recursion strategy in two-disk and three-disk versions of the Tower of Hanoi puzzle. Finally, all subjects were given Tower of Hanoi problems of increasing size, from three disks to eight disks, although several subjects were unable to complete the eight-disk puzzle; therefore, the data analysis concerns only the three-disk to seven-disk problems. The total number of moves required to solve a puzzle with N disks, using goal recursion, is $2^N - 1$. The start and goal pegs for puzzles of each size were selected at random from trial to trial. Between trials, subjects were reminded to use the goal-recursion strategy, and they were questioned at the end of all of the trials to ensure that they had complied. In place of a physical Tower of Hanoi, subjects saw a computer-generated (Vaxstation II) graphic display, with disks that moved when a source and destination peg were indicated with a mouse. Subjects seldom attempted an illegal move (placing a larger disk on a smaller disk), and such attempts were disallowed by the program. If subjects made a move that was inconsistent with the goal-recursion strategy (and hence would not move them toward the goal), the move was signaled as an error by a computer tone, and the subject was instructed to undo the erroneous move before making the next move. Thus, subjects could not stray more than one move from the optimal solution path. The main dependent measure was the total number of errors, that is, moves that were inconsistent with the goal-recursion strategy.

Subjects. The subjects were 45 students from Carnegie Mellon, the University of Pittsburgh, and the Community College of Allegheny County who participated for $10 payment. They took the Raven Advanced Progressive Matrices Test and solved the Tower of Hanoi puzzles.

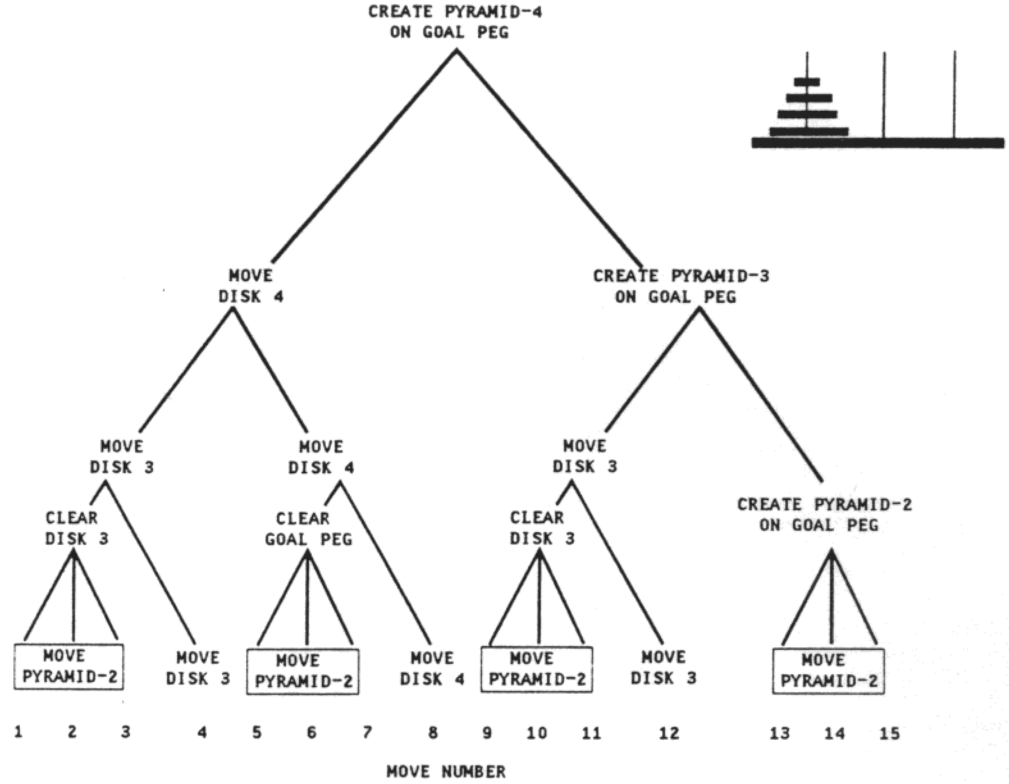

Figure 7. The goal tree generated by the goal-recursion strategy for the four-disk Tower of Hanoi puzzle. (The tree is traversed depth first, from left to right, generating the 15 moves.)

Results and Discussion

Because of its extensive dependence on goal management, overall performance of the goal-recursion strategy in the Tower of Hanoi puzzle was predicted to correlate highly with the Raven test. Consistent with this hypothesis, the correlation between errors on the Raven test and total number of errors on the six Tower of Hanoi puzzles was $r(43) = .77, p < .01$, a correlation that is close to the test–retest reliability typically found for the Raven test (Court & Raven, 1982). A subanalysis of the higher scoring subjects was also performed because many analyses presented later in this article deal primarily with students who scored in the upper half of our college sample on the Raven test. The subanalysis was restricted to subjects whose Raven scores were within 1 *SD* of the mean Raven score in Experiment 1a or above, eliminating 9 low-scoring subjects (scores between 12 and 17 points on the Raven test).[5] Even with this restricted range, the correlation between errors on the Tower of Hanoi puzzles and the Raven test for the 34 students with scores of 20 or higher was highly significant, $r(32) = .57$. These correlations support the thesis that the execution of the goal-recursion strategy in the Tower of Hanoi puzzle and performance on the Raven test are both related to the ability to generate and maintain goals in working memory.

A more specific prediction of the theory is that errors on the Tower of Hanoi puzzle should occur on moves that impose a greater burden on working memory and that the effect should depend, in part, on the capacity to maintain goals in working memory, as assessed by the Raven test. These predictions were supported, as shown in Figure 8, which shows the probability of an error on moves that require the generation of zero, one, or two or more subgoals; the four curves are for subjects who are classified according to their Raven test score. The error rates were low and comparable for moves that did not require the generation of additional subgoals; by contrast, lower scoring subjects made significantly more errors as the number of subgoals to be generated increased, as reflected in an interaction between the subject groups and whether there were zero or one

[5] As expected, subjects with low Raven test scores (12–17 points) made more errors than other groups as the number of subgoals to be generated increased (their error rate was .13, .66, and .59 for moves involving the generation of zero, one, and two or more subgoals, respectively). Their data were better fit by a model that assumed a more limited-capacity working memory and more goal generation, even for the smallest subpyramid. Also, the lowest scoring subjects were more likely to make multiple errors at a single move than were other subject groups. The lowest scoring subjects made an average of 17.2 of such errors while solving the five puzzles, compared with only 4 such errors by the next lowest scoring group (those with Raven test scores of 20–24 points). Multiple errors at a move suggest considerable difficulty in executing the strategy because only 2 errors were possible at each move, and 1 of the 2 consisted of retracting the previous correct move. Later in the article we discuss evidence that subjects in the bottom half of the distribution are more influenced than those in the upper half by extraneous processes while performing the Raven test as well.

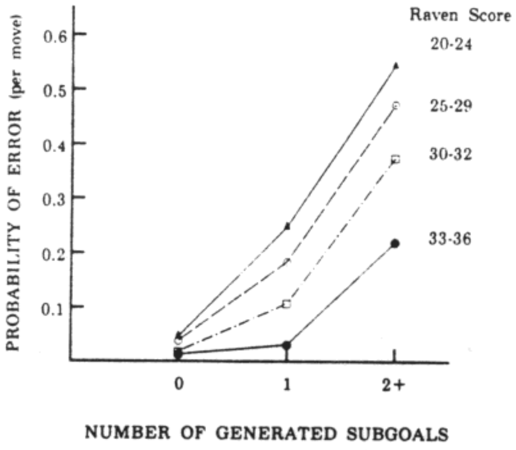

Figure 8. The probability of an error for moves in the Tower of Hanoi puzzle as a function of the number of subgoals that are generated to enable that move. (The curves represent subjects in Experiment 2 sorted according to their Raven test scores, from best [33–36 points] to low-median [20–25 points] performance.)

or more subgoals to be generated, $F(3, 32) = 3.57$, $p < .05$. Figure 8 also shows that the best performance was obtained by subjects with the best Raven test performance, $F(3, 32) = 3.53$, $p < .05$, and that the probability of an error increases with the number of subgoals to be generated in working memory, $F(2, 64) = 77.04$, $p < .01$. This pattern of results supports the hypothesis that errors in the Tower of Hanoi puzzle reflect the constraints of working memory: consequently, its correlation with the Raven test supports the theory that the Raven test also reflects the ability to generate and maintain goals in working memory.

The high correlation between the two tasks accounts for most of the reliable variance in the Raven test, raising the question of whether there is any need to postulate other processes, such as abstraction, as additional sources of individual differences in the Raven test. But using goal recursion in the Tower of Hanoi involves some abstraction to recognize each of the many configurations of subpyramids to which the strategy should be applied. Thus, the high correlation may reflect some shared abstraction processes, as well as goal generation and management.

The Raven test correlates with other cognitive tests that differ from it in form and content but, like the Raven test, appear to require considerable goal management. One example of such a test is an alphanumeric series-completion test, which requires the subject to determine which letter or number should occur next in a series, as in *1 B 3 D 5 G 7 K ? ?.* (The answer is *9 P.*)

Such correlations may reflect the fact that both tasks involve considerable goal generation and management. A theoretical analysis of the series-completion task by Kotovsky and Simon (1973; Simon & Kotovsky, 1963; Williams, 1972) indicated that the series-completion test, like the Raven test, requires correspondence finding, pairwise comparison of adjacent corresponding elements, and the induction of rules based on patterns of pairwise similarities and differences. The general similarity of the underlying processes leads to the prediction of correlated performance in the two tasks despite the minimal visuospatial pattern analysis in the series-completion task. This construal of the correlation is further supported by the fact that some of the sources of individual differences in the series-completion task are known and converge with our analysis of individual differences in the Raven test. Applying the Simon and Kotovsky (1963; Kotovsky & Simon, 1973) model to analyze the working memory load imposed by different types of series-completion problems, it was found that problems involving larger working memory loads differentiated between bright and average-IQ children more than did easier problems; this difference suggests that the ability to handle larger memory loads in the series-completion task correlates with IQ (Holzman, Pellegrino, & Glaser, 1983). These correlations, as well as the correlation between the Raven test and the Tower of Hanoi puzzle, strongly suggest that a major source of individual differences in the Raven test derives from the generation and maintenance of goals in working memory.

The Simulation Models

In this section, we first describe the FAIRAVEN model, which performs comparably to the median college student in our sample, who is already at a rather high level of performance relative to the population norms. Then, we describe the changes required to improve FAIRAVEN's performance to the highest level attained by our subjects, as instantiated by the BETTERAVEN model.

Overview

The primary goal in developing the simulation models was to specify the processes required to solve the Raven problems. In particular, the simulations should make explicit what distinguishes easier problems from harder problems, and correspondingly, what distinguishes among individuals of different ability levels. The simulations were designed to perform in a manner indicated by the performance characteristics observed in Experiment 1a, namely incremental, reiterative representation and rule induction.

The general outline of how the model should perform is as follows. The model encodes some of the figures in the first row of entries, starting with the first pair of entries. The attributes of the corresponding figures are compared, the remaining entry is encoded and compared with one of the other entries, and then the pattern of similarities and differences that emerges from the pairwise comparisons is recognized as an instance of a rule. In problems involving more than one rule, the model must determine which figural elements are governed by a common rule. The representation is constructed incrementally, and the rules are induced one by one. This process continues until a set of rules has been induced that is sufficient to account for all the variation among the entries in the top row. The second row is processed similarly, and in addition, a mapping is found between the rules for the second row and their counterparts in the first row. The rules for the top two rows are expressed in a generalized form and applied to the third row to generate the figural elements of the missing entry, and the generated missing entry is selected from the response alternatives.

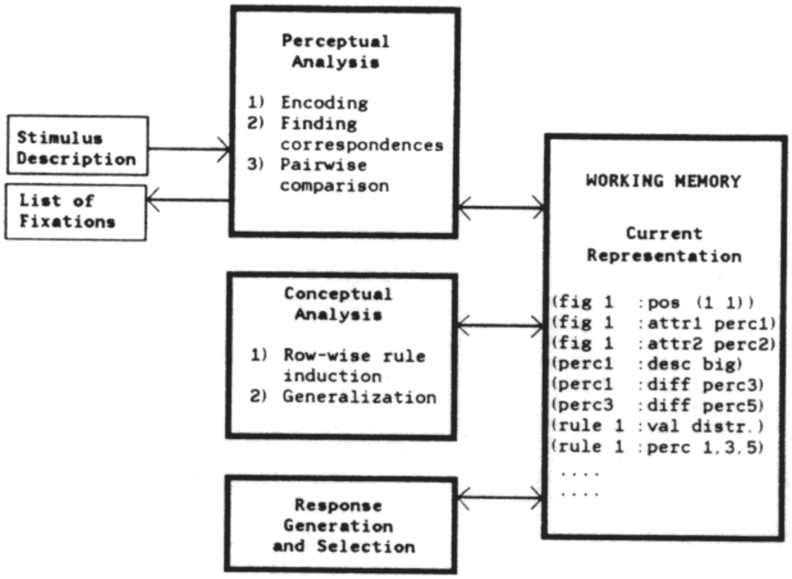

Figure 9. A block diagram of FAIRAVEN. (The perceptual-analysis productions, conceptual-analysis productions, and response-generation productions all interact through the contents of working memory. The perceptual-analysis productions accept stimulus descriptions and generate a list of simulated fixations. fig = figure; pos = position; attr = attribute; perc = percept; desc = description; diff = different; val = value; distr = distribution.)

The Programming Architecture

Both FAIRAVEN and BETTERAVEN are written as production systems, a formalism that was first used for psychological modeling by Newell and Simon (Newell, 1973; Newell & Simon, 1972) and their colleagues. In a production system, procedural knowledge is contained in modular units called *productions,* each of which specifies what actions are to be taken when a given set of conditions arises in working memory. If the conditions of one or more productions match the current contents of working memory, the productions are enabled to execute their actions, and they thereby change the contents of working memory (by modifying or adding to the contents). The new status of working memory then enables another set of productions, and so another cycle of processing starts. All production systems share these control principles, although they may differ along many other dimensions (see Klahr, Langley, & Neches, 1987).

The particular production-system architecture used for these simulations is CAPS (for Concurrent, Activation-Based Production System; Just & Carpenter, 1987; Just & Thibadeau, 1984; Thibadeau, Just, & Carpenter, 1982). Even though CAPS was constructed on top of a conventional system, OPS4 (Forgy & McDermott, 1977), it deviates in several ways from conventional production systems. One distinguishing property is that on any given cycle, CAPS permits all the productions whose conditions are satisfied to be enabled in parallel with each other. Thus, CAPS has the added capability of parallelism, along with the inherent seriality of a production system. By contrast, conventional production systems enable only one production per cycle, regardless of how many of them have had their conditions met, requiring some method for arbitrating among satisfied productions. Another distinguishing property of CAPS is that knowledge elements can have varying degrees of activation, whereas in conventional systems, elements are either present or absent from working memory. Other properties of CAPS, not used in the present applications, are described elsewhere (Just & Thibadeau, 1984; Thibadeau et al., 1982).

FAIRAVEN

The FAIRAVEN model consists of 121 productions that can be roughly divided into three categories: perceptual analysis, conceptual analysis, and responding. These three categories, which respectively account for approximately 48%, 40%, and 12% of all the productions, are indicated in the block diagram in Figure 9. The productions that constitute the perceptual analyzer simulate some aspects of the visual inspection of the stimulus. These productions access information about the visual display from a stimulus description file and bring this information into working memory as percepts. These productions also notice some relations among percepts. The productions in the conceptual analyzer try to account for the variation among the entries in one or more rows by inducing rules that relate the entries. The responder uses the induced rules to generate a hypothesis about what the missing matrix entry should be, and it then determines which of the eight response alternatives best fits that hypothesis. The next sections describe each of the three categories in more detail. This description is followed by an example of how FAIRAVEN solved the problem shown in Figure 2.

Perceptual Analysis

The FAIRAVEN model operates on a stimulus description that consists of a hand-coded, symbolic description of each matrix entry. Thus, the visual encoding processes that generate the symbolic representation lie outside the scope of the model. This incompleteness does not compromise our analysis of individual differences, for three reasons. First, the high correlations between the Raven test and other nonvisual tests (such as letter-series completion and verbal analogies, shown in Figure 1, bottom panel) indicate that visual encoding processes are not a major source of individual differences. Second, our protocol studies of the Raven test suggest that subjects have no difficulty perceiving and encoding the figures in each entry of a problem, such as squares, lines, angles, and so on. Third, the protocols indicate that the subjects have difficulty determining the correspondences among figures and their attributes, a process that lies within the scope of the model.

Stimulus descriptions. The perceptual-analysis productions operate on a symbolic description of each matrix entry and response alternative. To generate these descriptions, an independent group of subjects was asked to describe the entries in each problem, one entry at a time, without any problem-solving goal. The modal verbal descriptions served as the basis for the stimulus descriptions. The typical descriptions were in terms of basic-level figures (e.g., a square or a line) and their attributes (e.g., striped: Rosch, 1975). For example, the entry in the upper left of the matrix shown in Figure 2 would be described as a concatenation of two figures, a diamond and a line, with the line having the attributes of orientation (vertical) and texture (dark). The stimulus description of some figures contained an additional level of detail that was accessed if the base-level description was insufficient to establish correspondences, as in the case of embedded figures.

The perceptual analysis is done by three subgroups of productions that (a) encode the information about the figures, (b) determine the correspondences, and (c) compare the figures in adjacent entries to obtain a pattern of pairwise similarities and differences. Each subgroup is described in turn.

Encoding productions. These productions, the only access path to the stimulus information, transfer some or all of the information from the description file into working memory when such information is requested. If the entries in a given problem contain figures with several attributes, then FAIRAVEN will go through multiple cycles of perceptual analysis of the entries in a row until all the attributes have been analyzed. This behavior of the model was intended to express the incremental processing and reiterative scanning of the entries that was evident in the human eye-fixation patterns. Some of the simulated inspections of the stimulus, like the initial inspection of an entry, are data driven. If an entry's position in the matrix is specified, one of the encoding productions returns the names of each figure in that entry and the number of figures, but not any attribute information. Other inspections can be driven by a specific conceptual goal, such as the need to determine attributes of a particular figure. If an entry's position and the name of a figure are specified, one of the encoding productions returns an attribute of the figure and, if requested, its value. These encoding productions, which are more conceptually driven, are evoked after hypotheses are formulated in the course of inducing and verifying rules.

Finding correspondences between figures. In most problems, because more than one rule is operating, it is necessary to conceptually group the figures in a row that are operated on by each rule. The main heuristic procedure that subjects seem to use is to hypothesize that figures having the same name (e.g., *line*) should be grouped together. Similarly, FAIRAVEN uses a matching-names heuristic, which hypothesizes that figures having the same name correspond to each other. A second heuristic rule used by FAIRAVEN is the matching-leftovers heuristic, which hypothesizes that if all but one of the figures (or attributes) in two adjacent entries have been grouped, then those leftover figures (or attributes) correspond to each other. For example, for the problem depicted in Figure 2, the matching-names heuristic hypothesizes the correspondence among the three lines, and the matching-leftovers heuristic hypothesizes correspondence among the geometric figures that are leftover in each entry.

The FAIRAVEN model also tries to establish correspondences between the figures in different rows by expressing how the rules from a previous row account for the variation in the new row, usually by generalizing the rule.

Pairwise comparison. The pairwise comparison productions perform the fundamental perceptual comparisons between figures or attributes that are hypothesized to correspond to each other and thus provide the data base for the conceptual processing. These productions determine whether the elements are the same or different with respect to one of their attributes. For example, consider a row of three entries consisting of successive sets of circles: o oo ooo. By comparing the circle in the first entry with the two circles in the second entry, these productions would establish that they differ in the attribute of numerosity, such that the second entry has one more circle. These productions would then determine that this difference also characterizes the relation between the second and third entries. Both of these differences would be noted in working memory and would serve as the input to a production that hypothesizes a systematic variation in the numerosity of the circles across the three columns. The human counterpart of the pairwise comparison processes may be responsible for the one or more pairs of gazes between two related entries in the eye-fixation protocols.

Conceptual Analysis

The conceptual-analysis productions induce the rules that account for the variation among the figures and attributes in each of the first two rows. For example, if the numerosity of an element is one in column 1, two in column 2, and three in column 3, then a rule-induction production would hypothesize that the variation in numerosity is governed by a rule that says "add one as you progress rightward from column to column." These are the types of rules FAIRAVEN knows:

- constant in a row
- quantitative pairwise progression
- distribution of three values
- figure addition or subtraction.

Note that this list of rules does not include the distribution-of-two-values rule even though it is one of the rules governing the variation in some of the problems. The reason for omitting this rule is that problems containing this rule could not be solved with FAIRAVEN's limited correspondence-finding ability.

Also, problems containing this rule were often unsolved by the median subjects whom FAIRAVEN was intended to simulate.[6]

The main information on which the rule-induction productions operate are the patterns of pairwise similarities and differences. When a particular pattern of variation in the entries has been encoded in working memory, it directly evokes the appropriate rule-inducing production. Some of the productions in this module induce a rule to account for just one row at a time, whereas others induce a generalized form of the rule by combining the rules that apply to corresponding figures in both the first and the second rows. The generalization is made by expressing the rules in terms of variables rather than using the actual values encountered in the first two rows. The more general form of the rules induced by the model are intended to be counterparts of the human subjects' verbal statements of the rules. In a later section, the simulation's and human subjects' statements of rules are compared with respect to their content and the time in the trial at which they occur.

The perceptual analysis and the conceptual analysis are applied to the second row much as to the first row, except that the processing of the second row includes one additional step, namely establishing correspondences between the figures in the first and second rows. The perceptual analysis of the first two entries in the third row is similar to the analysis of the second row, including encoding, finding correspondences, and doing pairwise comparisons to determine which figures or values vary and which are constant in the first two entries. When this processing has been done, the response-generation productions take over.

Response Generation and Selection

The productions in this module use the hypothesized rules and the information in the first two columns of the third row to generate the missing entry in the third column. The general form of the rule that applies to the first two rows must be instantiated in terms of the specific values encountered in the first two entries in the third row. In problems containing more than one rule, the interrow correspondence between figures indicates which rules to associate with which figures. The instantiated rule (or rules) is applied to generate the missing entry, and then FAIRAVEN searches through the response alternatives for one that adequately matches the generated missing entry.

The model's strategy of generating the figures and attributes of the missing entry and then finding it among the alternatives closely corresponds to what the higher scoring subjects did. The lower scoring subjects sometimes scanned the response alternatives before inducing the rules, particularly in the case of the more difficult problems. Other researchers have also found that lower scoring subjects are more likely to use response-elimination strategies for geometric analogy problems, whereas higher scoring subjects are more likely to determine the properties of the desired response before examining the response alternatives (Bethell-Fox, Lohman, & Snow, 1984; Dillon & Stevenson-Hicks, 1981).

An Example of FAIRAVEN's Performance

The processes of FAIRAVEN can be illustrated by describing how it solves the problem depicted in Figure 2. The model starts by examining the top row. The variation among the three entries in a row is found by examining the pairwise similarities and differences between the figures and attributes found in adjacent columns. The first pairwise comparison is between the entries in the first and second columns of the top row. The encoding productions determine that the first entry contains a diamond and line and the second entry contains a square and line. The productions that find correspondences use the matching-names heuristic to postulate a correspondence between the lines that occur in the two entries. Once a correspondence is found between the lines, the matching-leftovers heuristic is used to postulate a second correspondence between the diamond and the square. Then FAIRAVEN compares the entries in the second and third columns. The lines in the second and third columns are postulated to correspond to each other, and the square is postulated to correspond to the triangle. The pattern of variation among the lines evokes the induction of a rule requiring that each entry in a row contain a line. Note that this is not the final form of the rule. The pattern of variation among the other figures in correspondence, namely the diamond, square, and triangle, evokes the induction of a distribution-of-three-values rule, such that each row contains one each of a diamond, square, and triangle in its three entries.

After these two rules have been induced, there is a second iteration of inspecting the entries in the first row. In the second iteration, the variation in the texture of the lines is noted, and this evokes the rule that each set of lines in a row has a texture that is either black, striped, or clear. On this and subsequent iterations, one attribute (and its value) per figure is perceived. Thus, the total number of iterations on a row depends on the maximum number of attributes possessed by any of the figures. As the variation in each additional attribute is discovered, one or more additional rules are induced to account for the variation. Thus, the perceptual and conceptual analyses are temporally interwoven.

The order in which the various attributes are processed is determined by the order in which they are encoded, which in turn is determined by their order in the stimulus description file, which in turn was guided by their order of mention by the subjects who only described the entries. So on the next iteration, FAIRAVEN encodes the orientation of the line. The value is vertical for each line, so FAIRAVEN hypothesizes a constant-in-a-row rule. The final (null) iteration reveals no further percepts to be accounted for, so FAIRAVEN proceeds to the second row. Note the similarity of FAIRAVEN's processing to the protocol of the human subject shown in Table 2, reflecting the incremental reiterative nature of the processing. For both the model and the subject, there are multiple visual scans of the first row, and in

[6] This list also omits another rule, so obvious as to be overlooked in our task analyses and the subjects' verbal reports, but not by the simulation model. The overlooked rule is the constant-everywhere rule. An example of this rule can be found in the problem shown in Figure 4b, in which every entry contains a diamond outline. In this particular example, the constant-everywhere rule does not discriminate among the response alternatives because they all contain a diamond outline. Both FAIRAVEN and BETTERAVEN used a constant-everywhere rule where applicable, but we do not discuss it further because of the minor role it plays in problem difficulty and individual differences.

both cases, there is a considerable time interval between the induction of the different rules.

The processing of the second row closely resembles that of the first row, in that the lines and geometric shapes are encoded and the correspondence among the lines and among the shapes is noticed. The rules governing the geometric shapes, line textures, and line orientation are induced. In addition, the correspondences between the geometric shapes in the first two rows is noticed, as is the correspondence between the lines, and a mapping is made between the rules for the two rows. The model notes that the rules governing line orientation are different in the two rows (constant vertical orientation in the first row and horizontal in the second row). Note that the subject's eye-fixation protocol in Table 2 shows a scan of row 2 interspersed with scattered inspections of row 1, which may reflect the mappings from one row to another.

The model proceeds to row 3, having formulated a generalized form of the rules, namely distribution of the three geometric shapes, distribution of the three line textures, and a constant orientation of lines in all the entries in a row. The inspection of the geometric figures in the first two columns of row 3 indicates which one of the triplet of shapes is missing (the square). Inspection of the line textures indicates which is missing (the black). Finally, the orientation of the lines in the first two columns indicates that the constant value of line orientation will be slanted from upper left to lower right.

The application of the three rules to the knowledge about the first two entries of row 3 is sufficient to correctly generate the missing entry, a square and a line. Only the three response alternatives that contain a square and line (2, 5, and 8) are considered further. The generated missing entry contains a black slanted (from upper left to lower right) line that matches Alternative 5, which is chosen as the answer.

The FAIRAVEN model solved 23 of the 34 problems it was given, the same as the median score of the 12 Carnegie Mellon students in Experiment 1a. Like the median subjects it was intended to simulate, FAIRAVEN solved the easier problems and could not solve most of the harder problems. The point-biserial correlation between the error rate for each problem in Experiment 1a and a dichotomous coding of FAIRAVEN's success or failure on the problem was $r(32) = .67, p < .01$, indicating that the model was more likely to succeed on the same problems that were solved by more of the human subjects. The performance of FAIRAVEN on each problem is given in the Appendix. However, we postpone a detailed analysis of the errors to the next section.

For the present purposes, the important point is that FAIRAVEN performed credibly, but at the same time, it had several limitations that prevented it from solving more problems. First, FAIRAVEN had no ability to induce rules that do not contain corresponding arguments (figures or attributes) in all three columns. Consequently, FAIRAVEN could not solve the problems involving the distribution-of-two-values rule. Second, FAIRAVEN had difficulty with problems in which the correspondence among figural elements is not discovered by either the matching-names heuristic or the matching-leftovers heuristic, such as correspondences based on location or texture. In these cases, the initially hypothesized correspondences based on figure names did not result in a correct rule, but FAIRAVEN had no way to backtrack. Third, FAIRAVEN had difficulty when too many high-level goals arose at the same time, and FAIRAVEN tried to pursue them concurrently. This situation occurred in problems with three or four rule tokens, when the perceptual evidence to support the multiple rules emerged simultaneously. The model tried to confirm all the rules in parallel, as CAPS permits, but the resulting bookkeeping load was unmanageable. In spite of these limitations, this program was able to perform on an intelligence test as well as some college students, using strategies similar to theirs and exhibiting behavior similar to theirs.

BETTERAVEN

The higher scoring subjects in our experiments performed better than FAIRAVEN; what psychological processes distinguish them from the median-scoring subjects and from FAIRAVEN? The BETTERAVEN model is our best current answer. The development of BETTERAVEN used FAIRAVEN as a starting point and did as little reorganization and addition as possible. The resulting model, BETTERAVEN, exercises more direct strategic control over its processes. Also, BETTERAVEN can induce more abstract rules on the basis of more abstract correspondences (permitting null arguments).

To improve BETTERAVEN's strategic control required the addition of a fourth category of productions, as shown in the block diagram in Figure 10. The new category is a goal monitor that sets strategic and tactical goals, monitors progress toward them, and adjusts the goals if necessary. In addition, the control structure of BETTERAVEN, as governed by the goal monitor, is somewhat changed. In BETTERAVEN, only one category of productions can be operating at a given time. The BETTERAVEN model also had some changes made to the perceptual and conceptual analyzers. The correspondence-finding processes are more sophisticated, allowing BETTERAVEN to handle rules applying to null arguments, such as a distribution-of-two-values rule. The conceptual analyzer also has more rules in its repertoire and uses the goal monitor to control the order in which rules are induced. The responder is effectively unchanged from FAIRAVEN.

The Goal Monitor

A module containing 15 productions sets main goals and subgoals for the model. The main purposes of the goal monitor are to ensure that higher level processes (namely rule induction) occur serially and not concurrently to provide an effective serial order for inducing rules (i.e., conflict resolution), to maintain an accounting of the model's progress toward its goals, and to appropriately modify its path to the solution when a difficulty is encountered. The goal monitor has a knowledge base that contains the goal structure for this task. For example, when starting to work a new problem, the goal monitor might set the following goals and subgoals and keep a record of their satisfaction or nonsatisfaction:

Top goal: Solve problem.
  Subgoal 1: Find all rules in top row.
    Subgoal 2: Do a first scan of top row.
      Subgoal 3: Compare adjacent entries.
        Subgoal 4: Find what aspects are the same or different or have no relation.

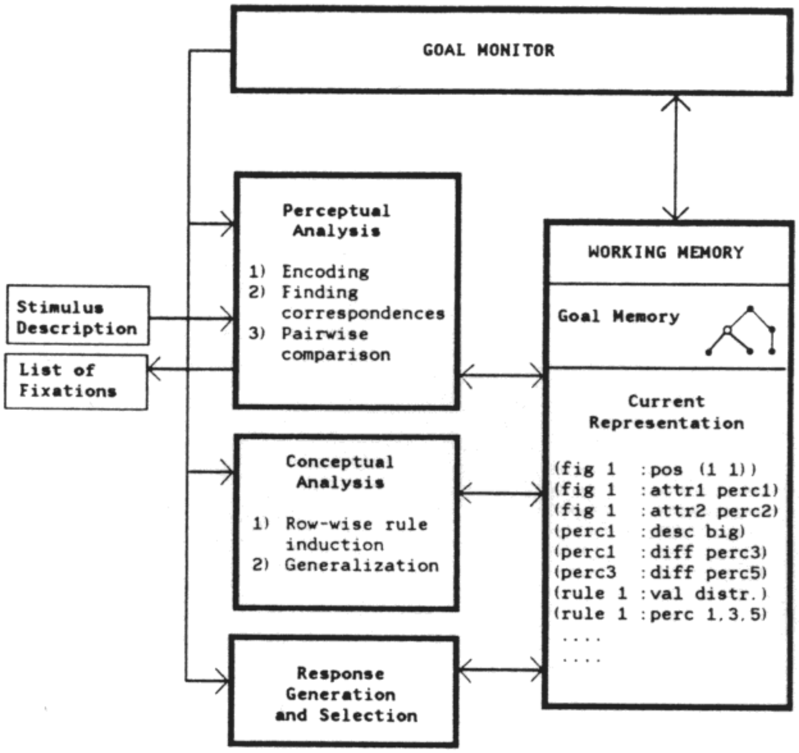

Figure 10. A block diagram of BETTERAVEN. (The distinction from FAIRAVEN visible from the block diagram is the inclusion of a goal monitor that generates and keeps track of progress in a goal tree. fig = figure; pos = position; attr = attribute; perc = percept; desc = description; diff = different; val = value; distr = distribution.)

To attain these goals, each row is reiteratively scanned, and rules are induced to account for the variation, with the number of iterations increasing with the complexity of the entries. This behavior of the model is motivated by the reiterative nature of the eye-fixation data and by the concurrent verbal protocols.

The management of the goal stack is under the exclusive control of the goal monitor. When it is appropriate to change the model's level of analysis, the goal monitor changes the current goal to either a parent goal or a subgoal. The consequence of setting a particular goal is to evoke some subset (module) of productions, such as the perceptual-analysis module or the response-generation module. The monitor keeps a record of goals that have been set and the current goal. This knowledge makes it possible to backtrack where necessary. Four backtracking productions take back specific hypothesized rules that have proved unfruitful, as well as taking back hypotheses about what the relevant attribute is and which elements correspond to each other. It is important to note that both BETTERAVEN and FAIRAVEN have goal-management capability, but BETTERAVEN's capability was enhanced as described.

Changes in the Perceptual Analyzer

The major change to the perceptual analyzer is that the heuristics for finding correspondences among figures are more general, overcoming several difficulties encountered by FAIRAVEN's heuristics. One type of difficulty arose when the number of figures per entry was not the same in each of the entries in a row. This difficulty occurs in problems containing a distribution-of-two-values rule, as well as figure addition and subtraction, in which a figure in one entry has no counterpart in another entry. Because FAIRAVEN assigned a counterpart to every figure in every entry, it would err in such problems (as did many of the lower scoring subjects). To deal with such rules, BETTERAVEN's new correspondence-finding productions in the perceptual analyzer assign a leftover element in one of a pair of entries to a null counterpart in another entry.

A second type of difficulty arose when the correspondence was based on an attribute other than the figure's name (such as two different figures having the same texture or position). When the matching-names (or any other) heuristic fails to lead to a satisfactory rule, BETTERAVEN's goal monitor can backtrack, postulate a correspondence based on an alternative attribute, and proceed thenceforth. By contrast, FAIRAVEN kept no record of choosing a correspondence heuristic and had no way of backing up to it if the choice turned out incorrect.

Rule Induction

Rule induction in BETTERAVEN was improved over FAIRAVEN's by virtue of serial rule induction (imposed by the goal monitor), the presence of a new rule (distribution of two values), and more general rules for figure addition and subtraction (enabled by the improved correspondence finding). Furthermore, the goal monitor permits BETTERAVEN to backtrack when a postulated rule fails to account for the variation.

In problems containing more difficult rules and a larger number of rules, FAIRAVEN's concurrent postulation of multiple rules led to several difficulties. First, there were competing attempts to simultaneously account for the same variation with two or more rules, which made the bookkeeping requirements unacceptably large in FAIRAVEN. Second, in problems with many figures, there was so much variation in figures that some of the arcane variation did not become evident unless the more mundane variation was first accounted for or in some sense removed. For example, in one of the problems containing two rules, it is much easier for the program (and human subjects) to induce a distribution-of-two-values rule after the other figures have been accounted for with a figure-addition rule. Finally, the human verbal protocols strongly suggested that the subjects attempted to fit only one rule at a time to the figures. Although CAPS permits parallelism at all levels, attempts at parallelism at FAIRAVEN's higher conceptual levels (namely rule induction) wreaked havoc, whereas parallelism at the lower perceptual levels caused no difficulty.

To improve BETTERAVEN's performance, the model was permitted to induce only one rule at a time, and furthermore, productions from the different modules were not permitted to fire concurrently. So the perceptual and conceptual modules of BETTERAVEN differ from each other in two respects: the time at which they dominate (early in the trial for the perceptual module vs. late for the conceptual) and whether they can tolerate concurrence (concurrence for the perceptual vs. seriality for the conceptual).

The seriality of rule induction and consequent processing in BETTERAVEN is enforced by conflict-resolution rules that arbitrate between any of the rule types that are hypothesized concurrently. The priorities prevent the later rules from firing until the earlier rules have had a chance to try to account for the variation. The priority among the rule types in the model is the following:

1. constant in a row
2. quantitative pairwise progression
3. distribution of three values
4. figure addition or subtraction
5. distribution of two values.

The design of BETTERAVEN required that there be an ordering, so that only one rule would be induced at a time, as it was in the human performance. However, BETTERAVEN's design did not dictate what that ordering should be. Three partial orderings were derived, largely on the basis of several empirical results and logical considerations. First, the priority accorded to the constant-in-a-row rule is based on the fact that it accounts for the most straightforward, null variation and is so relatively easy that it sometimes goes unmentioned in the human protocols. However, recall that the data do not eliminate the possibility that this rule can be induced in parallel with others, so the ordering of this rule type should not be overinterpreted. Second, figure addition or subtraction has priority over the distribution-of-two-values rule because it accounts for more figures in a row (each of the addends plus the sum, for a total of four figural components), whereas the distribution-of-two-values rule accounts for only two figural components. Finally, quantitative pairwise progression is given priority over the distribution-of-two-values rule by human subjects, as we learned from a study briefly described below.

Jan Maarten Schraagen performed a study in our laboratory that compared the relative time of mention of quantitative pairwise progression rules and that of distribution-of-two-values rules. To control for the possibility that the order in which rules are induced depends primarily on the relative salience of the figural components to which they apply, two isomorphs of each problem were constructed, differing in which rule applied to which figural elements. For example, in one isomorph, a quantitative pairwise progression rule might apply to the numerosity of lines, and a distribution-of-two-values rule might apply to some triangles. In the other isomorph, the quantitative pairwise progression rule would apply to the triangles, whereas the distribution-of-two-values rule would apply to the numerosity of lines. There were 86 observations (interpretable verbal protocols in correctly solved problems), and in 83% of these observations, the pairwise quantitative rule was induced before the distribution-of-two-values rule. This empirical finding confirms that at least part of the order in which the simulation induces the rules corresponds to the order in which people do.

Comparing Human Performance to the Theory

In this section, we compare the human performance to the simulation models for three types of performance measures: error patterns, the content of the rules that were induced, and online measures, specifically, patterns of eye fixations and verbal reports.

Error Patterns

As described earlier, FAIRAVEN solved 23 of the 34 problems, which is the median score of the 12 subjects in Experiment 1a. (Recall that only 32 of the problems were classifiable within our taxonomy.) The BETTERAVEN model lived up to its name, solving all but the two unclassifiable problems, similar to the performance of the best subject in Experiment 1a. Thus, the performance of FAIRAVEN and BETTERAVEN resembles the median and best performance, respectively. The patterns of human errors are analyzed in more detail later, to determine what characteristics of the problems are associated with the variation in error rates. After this analysis, a more detailed comparison is made between the human error patterns and those of the simulation models.

Interpretable patterns of errors emerge when the problems are grouped according to the properties in our taxonomy. The error rates of problems grouped this way are given in Table 3. The rows of the table correspond to different problem types that are distinguished by the type of rule involved, the number of different types of rules, and the total number of rules (of any type) and whether some of the problems in that group involved difficult correspondence finding. The error patterns in Experiments 1a and 1b are generally consistent with each other, even though the two experiments are not exact replications, because only half of the problems in Experiment 1b are from the Raven

Table 3
Error Rate (in Percent) for Different Problem Types

			Experiment	
No. rule types	No. rule tokens	Rule type	1a ($n = 12$)	1b ($n = 22$)
1	1	Pairwise progression	6	9
1	1	Addition or subtraction	17	13
1	1	Distribution of 3 values	29	25
1	2	Distribution of 3 values	29	21
2	2	Two different rules[a,b]	48	54
1	3, 4	Distribution of 2 values[b]	56	42
2	4	Distribution of 2 & 3 values[b]	59	54
1	3	Distribution of 3 values[b]	66	77

[a] This category is a miscellany of problems that contain two different rule types, such as addition and distribution of three values, or quantitative pairwise progression and distribution of three values.
[b] Corresponding elements are ambiguous or misleading for some or all of the problems in these categories.

Advanced Progressive Matrices Test and half are similar problems from the Standard Progressive Matrices Test.

In general, the error rates increase down the column as the number of rules in a problem increases. The lowest error rate, 6% in Experiment 1a and 9% in Experiment 1b, is associated with problems containing only a pairwise quantitative progression rule, indicating how easy this rule type was for our sample of subjects. Problems with pairwise quantitative progression rules may be relatively easy because, unlike all the other rules, this rule can be inferred from a pairwise comparison of only two figures. Repeated pairwise fixations between adjacent entries occurred frequently, even for lower scoring subjects. Pairwise comparison may be a basic building block of cognition, and consequently, it was made a basic architectural feature of the simulations.

The next lowest error rate is associated with problems that contain a single token of a figure addition or subtraction rule, or a distribution-of-three-values rule, shown in the second and third rows of Table 3. The rules relating the three entries in these problem types require that the subject consider all three arguments simultaneously, rather than only generalize one of the pairwise relations. To induce these types of rules, the subject must reason at a higher level of abstraction than that needed for pairwise similarities and differences. The verbal protocols in these problems indicated that the subjects who were having difficulty often persisted in searching for a single pairwise relation that accounted for the variation among all three entries.

The number of rule tokens appears to be a powerful determinant of error rate. The effect is seen clearly in the contrast between the relatively low error rate for problems with only one token (in the first three rows), averaging 16%, versus the error rate for problems with three or four tokens (in the last three rows), averaging 59%. One reason why inducing multiple rule tokens is harder is that it requires a greater number of iterations of rule induction to account for all of the variation. Moreover, keeping track of the variation associated with a first rule while inducing the second rule (or third rule) imposes an additional load on working memory. Approximately 50% of the errors on problems with multiple rules may arise from an incomplete analysis of the variation, as indicated in the ongoing verbal reports by a failure to mention at least one attribute or rule.[7] Such incompleteness may be partially attributed to failing to maintain the goal structures in working memory that keep track of what variation is accounted for and what variation remains unexplained. Another process made more difficult by multiple rules is correspondence finding. As the number of rules increases, so does the number of figural elements or the number of attributes that vary across a row. This, in turn, increases the difficulty of conceptually grouping the elements that are governed by each rule token.

The difficulties of correspondence finding were particularly apparent for problems with multiple possible correspondences and misleading cues to correspondences (like the problem in Figure 5 described earlier). An analysis of the subjects' verbal reports in all the problems identified as having misleading or ambiguous correspondence cues indicates that the correspondence-finding process was a source of significant difficulty. The reports accompanying 74% of the errors in these problems indicated that the subject had either postulated incorrect correspondences among figural elements or was not able to determine which elements corresponded. Sometimes subjects indicated this latter difficulty by saying that they could not see a pattern, even after extensive visual search or after having initially postulated and retracted various incorrect correspondences and rules.

In contrast to the types of rules listed in Table 3, the presence

[7] This estimate is based on problems containing either two or three tokens of the distribution-of-three-values rule. On 20% of the correct trials and 59% of the error trials, the verbal reports contained no evidence of the subject's having noticed at least one of the critical attributes: that is, neither the rule itself nor any attribute or value associated with that rule was mentioned. We assume that 20% is an estimate of how often the verbal reports do not reflect an encoded attribute. Consequently, we can estimate that 80% (the complement of 20%) of the 59% of the error trials with incomplete verbal reports (or approximately 50%) may be attributed to incomplete encoding or analysis, not just an omission in the verbal report.

of a constant-in-a-row rule had a small or negligible impact on performance. The mean error rate and response time for six problems containing the constant rule (involving distribution-of-three-values or figure addition or subtraction rules) were 30% and 38.9 s, respectively, which are similar to the measures for eight comparable problems that did not involve a constant rule (28% and 41.9 s). One possible reason for the minimal impact of a constant-in-a-row rule is that unlike any other type of rule, it requires storing only one value (i.e., the constant) for an attribute.

The analysis of the human error patterns can be compared with those of the simulation models. To make this comparison, the problems shown in Table 3 can be grouped further, dividing the table into the first four rows consisting of the easier problems and the last four rows consisting of the harder problems, namely those involving multiple rules, more abstract rules, misleading correspondences, or a combination of the three. The subjects to whom FAIRAVEN should be most similar are those with scores close to the median. The 6 subjects in Experiment 1a whose total score was within 10% of the median had a 17% error rate on the easier problems and a 70% error rate on the harder problems. In comparison, FAIRAVEN has a 0% error rate on the easier problems and a 90% error rate on the harder problems. Thus, the FAIRAVEN model has an error profile similar to the subjects it was intended to simulate, appropriately matching the difficulty these subjects have with problems containing multiple rule tokens and difficult correspondences. The BETTERAVEN model and the subjects to whom it should be similar (namely the best subjects) can solve almost all of the problems, so they have similar (essentially null) error profiles. The Appendix indicates the performance on each problem of Experiment 1a by the human subjects and by the two simulation models.

Modifications of BETTERAVEN

In addition to comparing FAIRAVEN and BETTERAVEN to the human performance, it is possible to degrade various abilities of BETTERAVEN and examine the resulting changes in performance. A demonstration that degraded versions of BETTERAVEN account for intermediate levels of performance between the levels of FAIRAVEN and BETTERAVEN can provide converging support for the present analysis of individual differences. Graceful degradation of BETTERAVEN also provides a sensitivity analysis that can indicate which of the new features of BETTERAVEN contributed to its improved performance. "Cognitive lesions" were made in BETTERAVEN to assess how its added features contributed to its superiority over FAIRAVEN. The two features of BETTERAVEN that were modified pertained to (a) abstraction, particularly the ability to induce the distribution-of-two-values rule, and (b) goal management.

Lesioning Abstraction Ability

One source of BETTERAVEN's advantage over FAIRAVEN is its ability to form abstract correspondences (involving null arguments) and hence induce the distribution-of-two-values rule. The BETTERAVEN model used this rule in 9 of the 11 most difficult problems; these were all problems that FAIRAVEN did not solve and BETTERAVEN did. Because the abstraction ability was firmly enmeshed with BETTERAVEN's processing, it was not possible to lesion it without disabling BETTERAVEN entirely. However, it was possible to lesion (eliminate) the distribution-of-two-values rule from BETTERAVEN's repertoire, in a model called *BETTERAVEN-without-distribution-of-2-rule*. Not surprisingly, this modified model did not correctly solve the 9 problems in which the rule had been used by BETTERAVEN (as shown in the Appendix), degrading its performance to the level of FAIRAVEN. However, it would be incorrect to conclude that this rule is the only property on which BETTERAVEN's superiority over FAIRAVEN is based, for two reasons. First, the ability to correctly induce the distribution-of-two-values rule depends on BETTERAVEN's ability to induce abstract correspondences, including the absence of an element. Second, this rule was evoked in problems involving multiple rules, and consequently, problems that taxed BETTERAVEN's goal management. As the next section demonstrates, the ability to manage goals also played a central role in BETTERAVEN's improvement over FAIRAVEN.

Lesioning Goal Management

To examine how BETTERAVEN's performance is influenced by goal-management capabilities, impaired versions of BETTERAVEN were created in which goal management competed with the ability to maintain and apply rules, to the extent that goal information was displaced from working memory. For example, in one of the lesioned models, if the problem required more than three rules to be induced and applied to the last row, then the extra rules (beyond three) displaced some of the remaining subgoals stored in the goal tree and resulted in an erroneous response in which only three rules were used to generate the response. This behavior corresponds to the human errors that are based on an incomplete set of rules. The modified versions of BETTERAVEN, which could maintain and apply either three, four, or five rules before displacing goals from working memory, are called *BETTERAVEN-3-rules*, *BETTERAVEN-4-rules*, and *BETTERAVEN-5-rules*, respectively. The performance of these modified versions is shown in the Appendix, along with the performance of the unmodified BETTERAVEN. In general, as the goal-management information in BETTERAVEN was increasingly displaced by information about the rules, its ability to solve problems was degraded. The BETTERAVEN-3-rules model solved 11 fewer problems than did the unmodified BETTERAVEN. The BETTERAVEN-4-rules model solved 8 fewer problems, and BETTERAVEN-5-rules solved 4 fewer problems than did the unmodified BETTERAVEN. The failures of the modified versions occurred primarily on problems with more rule tokens, namely the problems that require more goal management.

The cognitive lesioning experiments produced intermediate levels of performance, accounting for the continuum of performance that lies between FAIRAVEN and BETTERAVEN. Moreover, the relation between the particular lesions and the resulting patterns of errors confirms the importance of abstraction and goal management in performing the Raven test.

The Rules That Were Induced

The simulations can be evaluated in terms of the specific rules that they induce, in comparison with the rules of the subjects in Experiment 1b, who were instructed to try to solve each problem and then explicitly describe the rules they induced. The main comparison is based on rule descriptions provided by a plurality of the 12 (of 22) subjects modeled by FAIRAVEN and BETTERAVEN, namely those 12 who attained at least the median score. Across the 28 problems in Experiment 1b, there was a total of 59 attributes for which at least 1 subject gave a rule that was classifiable by our taxonomy.[8]

The main finding is that for 52 of the 59 attributes, BETTERAVEN induced the same rule as the plurality of the subjects. Four of the seven disagreements arose in cases where BETTERAVEN induced a distribution-of-two-values rule, whereas the subjects induced figure addition or subtraction.[9] The fit for FAIRAVEN was similar, except for problems involving the distribution-of-two-values rule, which FAIRAVEN did not solve. Thus, the simulation models match the subjects not only in which problems they solve but also in the rules that they induce.

In problems in which alternative rules can account for the same variation, there is a suggestion that higher scoring subjects induced different rules than did lower scoring subjects. Consider again the earlier example of how two different rules might describe a series of arrows pointing to 12, 4, and 8 o'clock; the variation can be described as distribution of three values or as a pairwise quantitative progression of an arrow's orientation, namely a clockwise rotation of 120°, beginning at 12 o'clock. Although both rules are sufficient to solve the problem, the transformational rule is preferable because it is typically more compact and generative; knowing the transformation and one of the values of an attribute is sufficient to generate the other two values (in the case of a quantitative progression rule), and the transformational rule usually applies more directly to successive rows. The verbal protocols were examined to determine whether transformational rules were more closely associated with correct solutions than distributional rules in the particular problems in which they were induced and whether more generally, higher scoring subjects were more likely to use transformational rules than distributional rules, compared with the lower scoring subjects.

Twenty-one of the 34 problems in Experiment 1a (Set I, Problem 12; Set II, Problems 1, 3, 8, 10, 13, 16, 17, 22, 23, 26, 27, 29, and 31–36) evoked a mixture of transformational and distributional rules from different subjects. Each protocol for these problems was categorized as describing a transformation or distribution of values, including partial descriptions. A description that had both transformational and distributional characteristics was counted as transformational. There were 156 transformational responses, 90 distributional responses, and 6 that could not be classified in Experiment 1a. For the problems in question, the transformational responses were associated with considerably better performance (error rate of 31%) than were the distributional responses (53% error rate). In a separate analysis limited to only those problems in which a correct final response was made, 71% of the problems were accompanied by a transformational rule, whereas 29% were accompanied only by a distributional rule. Transformational descriptions were not only associated with success in the problem in which they occurred, they were also associated with subjects who did well on the test as a whole. Higher scoring subjects were more likely to give transformational rules and less likely to give distributional rules. The ratio of transformational descriptions to distributional was 3.6:1 for the highest scoring subjects but only 1:1 for the lowest scoring subjects. These results associate transformational rules with better performance.

The rule ordering in BETTERAVEN (e.g., giving precedence of the pairwise quantitative progression rule over the distribution-of-three-values rule) provides a tentative account for the finding that a transformational rule was strongly associated with better performance. It is possible that only the higher scoring subjects have systematic preferences for some rule types over others, as BETTERAVEN does. By contrast, the choice among alternative rules may be random or in a different order for the lower scoring subjects. Thus, the differences in preferences among alternative rules between the higher and lower scoring subjects can be accommodated by an existing mechanism in BETTERAVEN.

On-Line Measures

The preliminary description of the results in Experiment 1a indicated that what was common to most of the problems and most of the subjects was the incremental problem solving. The incremental nature of the processing was evident in both the verbal reports and eye fixations. In problems containing more than one rule, the rules are described one at a time, with substantial intervals between rules. Also, the induction of each rule consists of many small steps, reflected in the pairwise comparison of related entries. We now examine the incremental processing in the human performance in more detail in light of the theoretical models, and we compare the human performance with the simulations' performance. The analyses focus on the effects of the number of rules in a problem on the number and timing of the reiterations of a behavior. To eliminate the effects of differences among types of rules, the analyses are limited to those problems that contained one, two, or three tokens of a distribution-of-three-values rule, and no other types of rules.

Inducing One Rule at a Time: Verbal Statements

The first way in which the rule induction is incremental is that in problems with multiple rules, only one rule is described

[8] The 12 subjects fully described a classifiable rule in only 51% of the 708 (12 × 59) cases. The agreement among subjects who described a rule for a given attribute was very high (only 7% of the 708 cases were disagreements with a plurality); however, in 37% of the cases, the rule descriptions were absent or incomplete, so the pluralities are sometimes small.

[9] The reason for BETTERAVEN's not using figure addition or subtraction in these cases is that they required a more general form of addition or subtraction than BETTERAVEN could handle. In one problem, the horizontal position of the figural element that was being subtracted was also being changed (operated on by another rule) from one column to the next. In the other problem, some figural elements had to be subtracted in one row but added in another row, so both types of variation would have to have been recognized as a form of a general figure addition or subtraction. Because BETTERAVEN's addition and subtraction rules were too specific to apply to these two situations, its distribution-of-two-values rule applied instead. The human subjects' ability to use an addition or subtraction rule testifies to the greater generality of their version of the rule compared with BETTERAVEN.

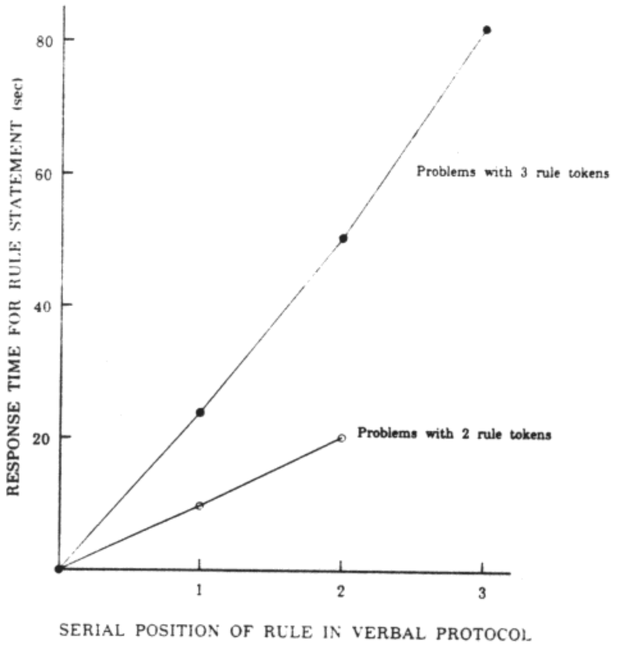

Figure 11. The elapsed time from the beginning of the trial to the verbal description of each of the rules in a problem.

at a time. The subjects appear to develop a description of one of the attributes in a row of entries, formulate it as a rule, verify whether it fits, and then go on to consider other unaccounted-for variation. This psychological process is a little like a stepwise regression, accounting for the variance with one rule, then returning to account for the remaining variance with another rule, and so on.

An analysis of the times at which the subjects report the rules in their verbal protocols strongly supports the interpretation that the rules are induced one at a time. In problems involving multiple rules, subjects generally stated each rule separately, with an approximately equal time interval separating the statements of the different rules. In the scoring of the verbal protocols, if a subject stated only the value of the attribute specified by the correct rule (e.g., "need a horizontal line") without stating the rule itself, this was counted as a statement of the rule. The descriptive statistic plotted in Figure 11 indicates the elapsed time from the beginning of the trial until the statement of the first rule, the second rule, and if there was one, the third rule. The verbal reports show a clear temporal separation between the statements of successive rules. The interval is much longer than the time needed to just verbalize the rules and seems most likely to reflect the fact that subjects induce the rules one at a time. The statement of a rule may lag behind the induction processes, but the long time between the rule statements strongly suggests that induction processes are serially executed. The time from the beginning of the trial until the first rule was stated was approximately 10 s for the five problems that had two rules per row; it then took another 10 s, on average, until the subject stated the second rule. Thus, it took an approximately similar amount of time to induce each of the rules. For the two more difficult problems, those involving three rules, the average time between each statement was close to 24 s. The fact that the interstatement times were longer for the latter group of problems indicates that a rule takes longer to induce if there is additional variation among the entries (variation that eventually was accounted for by the additional rules). Several of the processes would be made more difficult by the additional variation, particularly correspondence finding and goal management.

In contrast to the data just presented, which were based on 33 observations, there were four other trials in which subjects stated two rules together. In three of these cases, the time interval preceding the statement of the two rules together was approximately twice the time interval preceding the statement of single rules. We interpret this to mean that even when two rules are stated together, they may have still been induced serially, although we cannot rule out parallel processing of two rules at a slower rate on these four occasions.

The assertion that the rules are induced one at a time must be qualified to allow for the possibility that a constant-in-a-row rule might be processed on the same iteration as another rule. Most of the problems in the subset analyzed earlier contained a constant-in-a-row rule, but there was no systematic difference discernible in this small sample between problems that did or did not contain a constant rule. (Recall that a linear regression accounted for more of the variance among the mean error rates of problems if the count of rules excluded any constant rule.) Moreover, a constant-in-a-row rule was verbalized far less often than were the other types of rules. The structure of the stimulus set does not permit us to draw strong conclusions about the way the constant-in-a-row rule was processed.

The BETTERAVEN model is similar to the human subjects in inducing one rule at a time, in that there is a separation between the times at which the rules in a problem are induced. On average, there are 23 CAPS cycles (with range 22–24) between the time of inducing successive rule tokens. However, BETTERAVEN is unlike the students in several ways. First, the time between inducing rules is not affected by the number of rules (i.e., the amount of variation) in a problem: The 23-cycle interval applies equally to problems with two rule tokens and those with three rule tokens. By contrast, human subjects take longer to state a rule in problems with three rule tokens than in problems with two rule tokens, as shown in Figure 11. This difference suggests that BETTERAVEN's goal management is too efficient, relative to the human subjects. The model also differs from the students in its nonparallel induction of a constant rule (the 23-cycle time between rules disregards any induction of constant rules). In this respect, BETTERAVEN seems less efficient than the human subjects, who may be able to induce a constant rule in parallel with another rule.

Inducing One Rule at a Time: Eye-Fixation Patterns

Another way to demonstrate that the rules are induced one at a time is to compare the eye-fixation performance on problems containing increasing numbers of rules, looking for evidence of reiterations for problems with different numbers of rules. One of the most notable properties of the visual scan was its row-wise organization, consisting of repeated scans of the entries in a row. There was a strong tendency to begin with a scan of the top row and to proceed downward to horizontally scan each of the other two rows, with only occasional looks back to a pre-

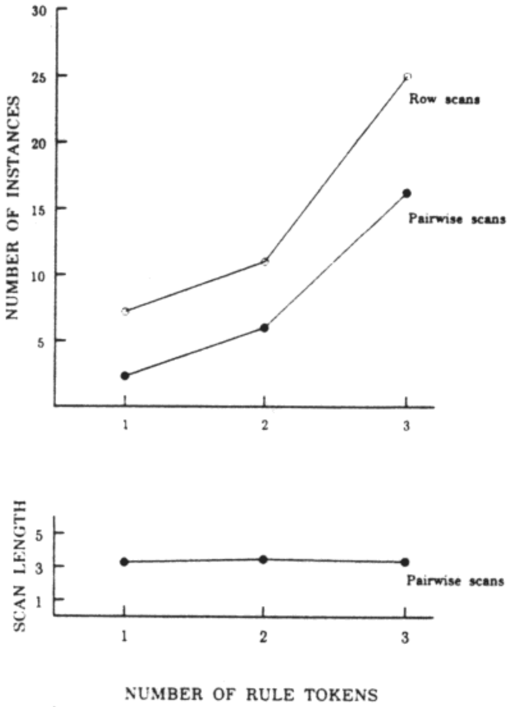

Figure 12. Top: The number of row scans and pairwise scans increases with the number of rule tokens in the problem. Bottom: Length of the pairwise scans (i.e., the number of alternating gazes between a pair of entries) is unaffected by the number of rule tokens.

viously scanned row. (This description applies particularly well to problems involving quantitative pairwise progression rules or addition or subtraction rules and slightly less well to the problems in the subset that is being analyzed here, involving multiple distribution-of-three-values rules. In the latter problems, subjects also used a row organization, but they sometimes looked back to previously scanned rows.) So it is reasonable to ask whether there is a dependence between the number of scans through the rows and the number of rules.

The data indicate that in general, the number of times that subjects visually scanned a row (or a column, or occasionally, a diagonal) increased with the number of rules in the problem. A scan of a row was defined as any uninterrupted sequence of gazes on all three of the entries in that row, allowing refixation of any of the entries (and was similarly defined for a column scan).[10] The analysis showed that as the number of rule tokens in a problem increases from one to two to three, the number of row scans increases from 7.2 to 11 to 25, as shown in the top panel of Figure 12. It is likely that the rule is being induced and verified during the multiple scans associated with each rule.

Incremental Processing in Inducing a Rule

There are many small steps in inducing each rule. For example, in a problem containing a quantitative pairwise progression rule, BETTERAVEN can induce the rule in a tentative form after a pairwise comparison between the entries in the first two columns in the row. Then the second and third columns can be compared, and a tentative rule can be induced, followed by a higher order comparison that verifies or disconfirms the correctness of the tentative rules. In the case of disconfirmations, all of the preceding processes must be reexecuted, generating additional pairwise comparisons. Thus, there are reiterative cycles of encoding stimulus properties, comparing properties between entries, inducing a rule, and verifying the rule's adequacy.

As the number of rules increases, so should the number of pairwise similarities and differences to be encoded and, consequently, the number of pairwise comparisons. The eye-fixation data provide clear evidence supporting this prediction. A pairwise scan was defined as any uninterrupted sequence of at least three gazes alternating between any two entries, excluding those that were part of a row or column scan because they had already been included in the row-scanning measures described earlier. Consistent with the theoretical prediction, as the number of rule tokens in a problem increased from one to two to three, the mean number of pairwise scans (of any length) increased from 2.3 instances to 6 to 16.2, as shown in the top panel of Figure 12.

We can also determine whether the difficulty of making a pairwise comparison (as indicated by the sequence length of a pairwise scan) also increases in the presence of additional variation between the entries (as indicated by the number of rules). As shown in the bottom panel of Figure 12, the number of rules in the problem had no effect on the mean sequence length of the pairwise scans. Thus, the pairwise scans may reflect some primitive comparison process that pertains to the induction of a single rule token and is uninfluenced by the presence of additional variation between the entries. This result is consistent with a theory that says that difficult problems are dealt with incrementally, by decomposing the solution into simple subprocesses. So some subprocesses, like the comparison of attributes of two elements, should remain simple in the face of complexity (Figure 12, bottom panel), even as other performance measures show complexity effects (Figure 12, top panel). The decomposition implied by the various forms of incremental processing observed here is probably a common way of dealing with complexity.

Limitations of the Model

At both the micro and macro levels, FAIRAVEN and BETTERAVEN perform comparably to the college students that they were intended to model. The models solve approximately the same

[10] The gaze analyses of the problems containing different numbers of tokens of distribution-of-three-values rules was applied to the protocols of 6 of 7 scorable subjects (who happened to be the higher scoring subjects), eliminating trials on which subjects made an error, or when the eye-fixation data were lost due to measurement noise. The seventh scorable eye-fixation protocol was excluded because it came from the lowest scoring subject, who had too few correct trials to contribute. The data in Figure 12 are from 10 observations of problems in Set I, Problem 7, and Set II, Problem 17, each of which contains one rule token: 25 observations of problems in Set I, Problems 8 and 9, and Set II, Problems 1, 13, and 27, which contain two rule tokens; and 5 observations of problems in Set II, Problems 29 and 34, which contain three rule tokens.

subsets of problems as the corresponding students, and they induce similar sets of rules. Also, the simulations resemble the students in their reliance on pairwise comparisons and in their sequential induction of the rules. The simulations are both sufficient and plausible descriptions of the organization of processes needed to solve these types of problems. The commonalities of the two programs, namely the incremental, reiterative processing, express some of the fundamental characteristics of problem solving. The differences between the programs, namely the nature of the goal management and abstraction, express the main differences among the individuals with respect to the processing tapped by this task. Although the simulations match the human data along many dimensions of performance, there are also differences. In this section, we address four such differences and their possible relation to individual differences in analytic intelligence.

Perhaps the most obvious difference between the simulations and the human performance is that the simulations lack the perceptual capabilities to visually encode the problems. However, as we argued earlier, this does not compromise our analysis of the nature of individual differences because numerous psychometric studies suggest that the visual encoding processes are not sources of individual differences in the Raven test. This is not to say that visual encoding and visual parsing processes do not contribute to the Raven test's difficulty, but only that such processes are not a primary source of individual differences. In addition, the success of the simulation models suggests that the strictly visual quality of the problems is not an important source of individual differences; analogous problems in other modalities containing haptic or verbal stimuli would be expected to similarly tax goal management and abstraction.

A second difference is that the simulations, unlike the students, do not read the instructions and organize their processes to solve the problems. Although this mobilization of processes is clearly an important part of the task and an important part of intelligence, it is an unlikely source of individual differences for this population. All of the college students could perform this task sufficiently well to solve the easier, quantitative pairwise comparison problems. Moreover, even though the metaprocesses that assemble and organize the processes lie outside the scope of the current simulation, they could be incorporated without fundamentally altering the programs or their architecture (e.g., see Williams, 1972).

A third feature that might appear to differentiate the simulations from human subjects is the difference between rule induction and rule recognition. Both FAIRAVEN and BETTERAVEN are given a set of possible rules, and they only have to recognize which ones are operating in a given problem, rather than inducing the rules from scratch. However, with the notable exception of the distribution-of-two-values rule, the other rules are common forms of variation that were correctly verbally described by all subjects in some problems. Hence, the individual differences were not in the knowledge of a particular rule so much as in recognizing it among other variation in problems with multiple rule tokens. By contrast, knowledge of the distribution-of-two-values rule appeared to be a source of individual differences. We account for its unique status in terms of its abstractness and unfamiliarity. In fact, we express the better abstraction capabilities of BETTERAVEN both in terms of its ability to handle a larger set of patterns of differences and its explicit knowledge of this rule. Thus, the difference between the two simulations expresses one sense in which knowledge of the rules distinguishes among individuals. On the other hand, BETTERAVEN does not have the generative capability of inducing all of the various types of abstract rules that one might encounter in these types of tasks; in this sense, it falls far short of representing the full repertoire of human induction abilities.

A fourth limitation of the models is that they are based on a sample of college students who represent the upper end of the distribution of Raven scores, and so the theoretical analysis cannot be assumed to generalize throughout the distribution. We would argue, however, that the characteristics that differentiate college students, namely goal management and abstraction, probably continue to characterize individual differences throughout the population. But there is also evidence that low-scoring subjects sometimes use very different processes on the Raven test, which could obscure the relationship between Raven test performance and working memory for such individuals. For example, as mentioned previously, low-scoring subjects rely more on a strategy of eliminating some of the response alternatives, fixating the alternatives much sooner than high-scoring subjects (Bethell-Fox et al., 1984; Dillon & Stevenson-Hicks, 1981). Moreover, the types of errors made by low-scoring adults frequently differ from those made by high-scoring subjects (Forbes, 1964) and may reflect less analysis of the problem.

If such extraneous processes are decreased and low-scoring subjects are trained to use the analytic strategies of high-scoring subjects, the validity of the Raven test increases. The study, with 425 Navy recruits, found that for low-scoring subjects, the correlation between the Raven test and a wide-ranging aptitude battery increased significantly (from .08 to .43) when the Raven problems were presented in a training program that was designed to reduce nonanalytic strategies (Larson, Alderton, & Kaupp, 1990). The training did not alter the correlation between the Raven test and the aptitude battery for subjects in the upper half of the distribution. The fact that the performance of the trained low-scoring and all of the high-scoring subjects correlated with the same aptitude battery suggests that after training, the Raven test drew on similar processes for each group. Thus, it is plausible to suppose that the current model could be generalized to account for the performance of subjects in the lower half of the distribution if they are given training to minimize the influence of extraneous processes.

Cognitive Processes and Human Intelligence

In this part of the article, we discuss the implications of the model for analytic intelligence. In the first sections, we examine how abstraction and goal management are realized in other cognitive tasks. These sections are focused primarily on goal management rather than abstraction, in part because abstraction is implicitly or explicitly incorporated into many theories of analytic intelligence, whereas goal management has received less attention. In the final section, we examine what the Raven simulations suggest about processes that are common across people and across different domains.

Abstraction

Most intuitive conceptions of intelligence include an ability to think abstractly, and certainly solving the Raven problems involves processes that deserve that label. Abstract reasoning consists of the construction of representations that are only loosely tied to perceptual inputs and instead are more dependent on high-level interpretations of inputs that provide a generalization over space and time. In the Raven test, more difficult problems tended to involve more abstract rules than the less difficult problems. (Interestingly, the level of abstraction of even the most difficult rule, distribution of two values, does not seem particularly great compared with the abstractions that are taught and acquired in various academic domains, such as physics or political science.) The level of abstraction also appears to differentiate the tests intended for children from those intended for adults. For example, one characterization of the easy problems found in the practice items of Set I and in the Coloured Progressive Matrices Test is that the solutions are closely tied to the perceptual format of the problem and consequently can be solved by perceptual processes (Hunt, 1974). By contrast, the problems that require analysis, including most of the problems in Set II of the Advanced Progressive Matrices Test, are not as closely tied to the perceptual format and require a more abstract characterization in terms of dimensions and attributes.

Abstract reasoning has been a component of most formal theories of intelligence, including those of traditional psychometricians, such as Thurstone (1938), and more recent researchers of individual differences (Sternberg, 1985). Also, Piaget's theory of intelligence characterizes childhood intellectual development as the progression from the concrete to the symbolic and abstract. We can now see precisely where the Raven test requires abstraction and how people differ in their ability to reason at various levels of abstraction in the Raven problems.

Goal Management

One of the main distinctions between higher scoring subjects and lower scoring subjects was the ability of the better subjects to successfully generate and manage their problem-solving goals in working memory. In this view, a key component of analytic intelligence is goal management, the process of spawning subgoals from goals, and then tracking the ensuing successful and unsuccessful pursuits of the subgoals on the path to satisfying higher level goals. Goal management enables the problem solver to construct a stable intermediate form of knowledge about his or her progress (Simon, 1969). In Simon's words, "complex systems will evolve from simple systems much more rapidly if there are stable intermediate forms than if there are not. The resulting complex forms in the former case will be hierarchic" (1969, p. 98). The creation and storage of subgoals and their interrelations permit a person to pursue tentative solution paths while preserving any previous progress. The decomposition of the complexity in the Raven test and many other problems consists of the recursive creation of solvable subproblems. The benefit of the decomposition is that an incremental iterative attack can then be applied to the simplified subproblems. A failure in one subgoal need not jeopardize previous subgoals that were successfully attained. Moreover, the record of failed subgoals minimizes fruitless reiteration along previously failed paths. But the cost of creating embedded subproblems, each with their own subgoals, is that they require the management of a hierarchy of goals.

Goal management probably interacts with another determinant of problem difficulty, namely the novelty of the problem. A novel task may require the organization of high-level goals, whereas the goals in a routine task have already been used to compile a set of procedures to satisfy them, and the behavior can be much more stimulus driven (Anderson, 1987). The use or organization of goals is a strategic level of thought, possibly involving metacognition or requiring reflection. In the BETTERAVEN model, additional goal-management mechanisms, such as selection among multiple goals, a goal monitor, and backup from goals, had to be included to solve the more difficult problems. However, if people had extensive practice or instruction on Raven problems, the goal management would become routine, thereby making the problems easier. Instruction of sixth graders in the use of the type of general strategy used by FAIRAVEN and BETTERAVEN improves their scores on Set I of the Raven test (Lawson & Kirby, 1981).

This analysis of the source of individual differences in the Raven test should apply to other complex cognitive tasks as well. The generality of the analysis is supported by Experiment 2, which showed a large correlation between the Raven test and the execution of a Tower of Hanoi puzzle strategy that places a large burden on goal generation and goal management. Our analysis is also consistent with the high correlations among complex reasoning tasks with diverse content, such as the data cited in the introduction (Marshalek et al., 1983; Snow et al., 1984). These researchers and others have suggested that the correlations among reasoning tasks may reflect higher level processes that are shared, such as executive assembly and control processes (see also Carroll, 1976; Sternberg, 1985). The contribution of the current analysis is to specify these higher level processes and instantiate them in the context of a widely used and complex psychometric test.

The Raven Test's Relation to Other Analogical Reasoning Tasks

The analogical nature of the Raven problems suggests that the Raven processing models should bear some resemblance to other models of analogical reasoning. One of the earliest such artificial intelligence projects was Evans's (1968) ANALOGY program, which solved geometric analogies of the form (A:B::C:[five choices]). Evans's program had three main steps. The program computed the spatial transformation that would transform A into B by using specific knowledge of analytic geometry. It then determined the transformation necessary to transform C into each of the five possible answers. Finally, it compared and identified which solution transformation was most similar to that for transforming A into B and returned the best choice. A major contribution of ANALOGY was that it specified the content of the relations and processes that were sufficient to solve problems from the American Council on Education examination. Although ANALOGY was not initially intended to account for human performance, Mulholland, Pellegrino, and Glaser (1980) found that aspects of the model accounted for the pattern of response times and errors in solving 2 × 2 geometric analogies. Errors and response times increased with the number of processing operations, which Mulholland et al. attributed to the increased burden on working memory incurred by tracking elements and transformations. Thus, much simpler analogical reasoning tasks can reflect working memory constraints.[11]

Analogical reasoning in the context of simple 2 × 2 matrices has also been analyzed from the perspective of individual differences. The theoretical issue has been whether individual differences in the speed of specific processes (e.g., inferencing, mapping, and verifying) account for individual differences in more complex induction tasks, like the Raven test. For example, Sternberg and Gardner (1983) found that a speed measure based on various inference processes used in simple analogical and induction tasks was correlated with psychometrically assessed reasoning ability. However, several other studies have failed to find significant correlations between the speed of specific inference processes and performance in a more complex reasoning task (Mulholland et al., 1980; Sternberg, 1977). The overall pattern of results suggests that the speed of any specific inference process is unlikely to be a major determinant of goal management. This conclusion is also supported by the high correlation between the Raven test and the Tower of Hanoi puzzle, a task that required very little induction. Nevertheless, some degree of efficiency in the more task-specific processes may be a necessary (if not sufficient) condition to free up working-memory resources for goal generation and management. The analysis of reasoning in simple analogies illuminates the task-specific inference processes but is unlikely to account for the individual differences in the more complex reasoning tasks.

What Aspects of Intelligence Are Common to Everyone?

The Raven test grew out of a scientific tradition that emphasizes the analysis of intelligence through the study of individual differences. The theoretical goal of the psychometric or differential approach (in contrast to its methodological reliance on factor analysis) is to account for individual performance, not simply some statistical average of group performance. The negative consequence of this approach is that it can conceptually and empirically exclude processes that are necessary for intelligent behavior, but are common to all people, and hence not the source of significant differences among individuals. Computational models such as the Raven simulations must include both the processes that are common across individuals and those that are sources of significant differences. Consequently, the models provide insights into some of the important aspects of intelligence, such as the incremental and reiterative nature of reasoning.

Cognitive accounts of other kinds of ability, such as models of spatial ability (e.g., Just & Carpenter, 1985; Kosslyn, 1980; Shepard & Cooper, 1982) and language ability (e.g., Just & Carpenter, 1987; van Dijk & Kintsch, 1983), also contribute to the characterizations of intelligence. Newell (in press) has argued that psychology is sufficiently mature to warrant the construction of unified theories of cognition that encompass all of the kinds of thinking included in intelligence and offers the SOAR model as his candidate. Although the collection of models for diverse tasks that we have developed is far more modest in scope, all of the models have been expressed in the same theoretical language (CAPS), making the commonalities and differences relatively discernible. All of these models share a production-system control structure, a capacity for both seriality and parallelism, a representational scheme that permits different activation levels, and an information accumulation function (effectively, an activation integrator). One interesting difference among tasks is that some types of processes are easy to simulate with parallelism and others are not (easy in the sense that the models can perform the task and still retain essential human performance characteristics). The processes that seem to operate well in parallel in the simulation models are highly practiced processes and lower level perceptual processes. The simulation of higher level conceptual processes is accomplished more easily with seriality, unless extensive increments to goal management are included.

What the theory postulates about the commonalities of different people and different tasks reflects some of the observed performance commonalities. Many of the performance commonalities occur at the microstructure of the processing, which is revealed in the eye-fixation patterns. The time scale of this analysis is approximately 300–700 ms per gaze. Such processes are too fast for awareness or for including in a verbal report. The eye-fixation analysis reveals iterations through small units of processing: the task is decomposed into manageable units of processing, each governed by a subgoal. Then, the subgoals are attacked one at a time. The problem decomposition and subgoaling reflect how people handle complexity beyond their existing operators in a number of domains, including text comprehension, spatial processing, and problem solving. For example, in a mental rotation task, subjects decomposed a cube into smaller units that they then rotated one unit at a time (Just & Carpenter, 1985). Similarly, in the Raven test, even the simplest types of figural analogies were decomposed and incrementally processed through a sequence of pairwise comparisons. This segmentation appears to be an inherent part of problem solving and a facet of thinking that is common across domains in various tasks requiring analytic intelligence.

Thus, what one intelligence test measures, according to the current theory, is the common ability to decompose problems into manageable segments and iterate through them, the differential ability to manage the hierarchy of goals and subgoals generated by this problem decomposition, and the differential ability to form higher level abstractions.

[11] Although there has been a large amount of subsequent artificial intelligence research on analogical reasoning, most of the work has focused on knowledge representation, knowledge retrieval, and knowledge utilization rather than inferential and computational processes (see the summary by Hall, 1989). Analogical reasoning is viewed as a bootstrapping process to promote learning and the application of old information in new domains (Becker, 1969; McDermott, 1979). In psychology, this view of analogical reasoning has resulted in research that examines the conditions under which subjects recognize analogous problem solutions (Gick & Holyoak, 1983) and the contribution of analogical reasoning to learning (Gentner, 1983).

References

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review, 94*, 192–210.

Becker, J. D. (1969). The modeling of simple analogic and inductive processes in a semantic memory system. In *Proceedings of the 1st International Joint Conference on Artificial Intelligence* (pp. 655–668). Bedford, MA: The MITRE Corporation.

Belmont, L., & Marolla, F. A. (1973). Birth order, family size, and intelligence. *Science, 182*, 1096–1101.

Bethell-Fox, C. E., Lohman, D., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence, 8*, 205–238.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect". In L. Resnick (Ed.), *The nature of intelligence* (pp. 27–56). Hillsdale, NJ: Erlbaum.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1–22.

Court, J. H., & Raven, J. (1982). *Manual for Raven's progressive matrices and vocabulary scales* (Research Supplement No. 2, Pt. 3, Section 7). London: H. K. Lewis.

Dillon, R. F., & Stevenson-Hicks, R. (1981). *Effects of item difficulty and method of test administration on eye scan patterns during analogical reasoning.* Unpublished Tech. Rep. No. 1, Southern Illinois University, Carbondale, Department of Psychology.

Egan, D. E., & Greeno, J. (1974). Theory of rule induction: Knowledge acquired in concept learning, serial pattern learning, and problem solving (pp. 43–103). In L. W. Gregg (Ed.), *Knowledge and cognition.* Hillsdale, NJ: Erlbaum.

Evans, T. G. (1968). A program for the solution of a class of geometric analogy intelligence test questions. In M. Minsky (Ed.), *Semantic information processing* (pp. 271–353). Cambridge, MA: MIT Press.

Forbes, A. R. (1964). An item analysis of the advanced matrices. *British Journal of Educational Psychology, 34*, 1–14.

Forgy, C., & McDermott, J. (1977). OPS: A domain-independent production system language. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence* (pp. 933–939). Pittsburgh, PA: Carnegie Mellon University, Department of Computer Science.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155–170.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38.

Hall, R. P. (1989). Computational approaches to analogical reasoning: A comparative analysis. *Artificial Intelligence, 39*, 39–120.

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

Hunt, E. B. (1974). Quote the Raven? Nevermore! In L. W. Gregg (Ed.), *Knowledge and cognition* (pp. 129–158). Hillsdale, NJ: Erlbaum.

Jacobs, P. I., & Vandeventer, M. (1972). Evaluating the teaching of intelligence. *Educational and Psychological Measurement, 32*, 235–248.

Jensen, A. R. (1987). The *g* beyond factor analysis. In R. R. Ronning, J. A. Glover, J. C. Conoley, & J. C. Witt (Eds.), *The influence of cognitive psychology on testing* (pp. 87–142). Hillsdale, NJ: Erlbaum.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology, 8*, 441–480.

Just, M. A., & Carpenter, P. A. (1979). The computer and eye processing pictures: The implementation of a raster graphics device. *Behavior Research Methods & Instrumentation, 11*, 172–176.

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review, 92*, 137–172.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension.* Newton, MA: Allyn & Bacon.

Just, M. A., & Thibadeau, R. H. (1984). Developing a computer model of reading times. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 349–364). Hillsdale, NJ: Erlbaum.

Klahr, D., Langley, P., & Neches, R. (Eds.). (1987). *Production system models of learning and development.* Cambridge, MA: MIT Press.

Kosslyn, S. M. (1980). *Image and mind.* Cambridge, MA: Harvard University Press.

Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology, 17*, 248–294.

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

Larson, G. E., Alderton, D. L., & Kaupp, M. A. (1990). *Construct validity of Raven's progressive matrices as a function of aptitude level and testing procedures.* Unpublished manuscript, Navy Personnel Research and Development Center, Testing Systems Department, San Diego, CA.

Lawson, M. J., & Kirby, J. R. (1981). Training in information processing algorithms. *British Journal of Educational Psychology, 51*, 321–335.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7*, 107–127.

McDermott, J. (1979). Learning to use analogies. In *Proceedings of the 6th International Joint Conference on Artificial Intelligence* (pp. 568–576). Tokyo: Information Processing Society of Japan.

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

Newell, A. (1973). Production system: Models of control structures. In W. G. Chase (Ed.), *Visual information processing* (pp. 463–526). New York: Academic Press.

Newell, A. (in press). *Unified theories of cognition: The 1987 William James Lectures.* Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Raven, J. C. (1962). *Advanced Progressive Matrices, Set II.* London: H. K. Lewis. (Distributed in the United States by The Psychological Corporation, San Antonio, TX)

Raven, J. C. (1965). *Advanced Progressive Matrices, Sets I and II.* London: H. K. Lewis. (Distributed in the United States by The Psychological Corporation, San Antonio, TX)

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192–233.

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations.* Cambridge, MA: MIT Press.

Simon, H. A. (1969). *The sciences of the artificial.* Cambridge, MA: MIT Press.

Simon, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology, 7*, 268–288.

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review, 70*, 534–546.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 47–103). Hillsdale, NJ: Erlbaum.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Sternberg, R. J. (1977). *Intelligence, information processing and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* New York: Cambridge University Press.

Sternberg, R. J., & Gardner, M. K. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General, 112,* 80–116.

Thibadeau, R., Just, M. A., & Carpenter, P. A. (1982). A model of the time course and content of reading. *Cognitive Science, 6,* 157–203.

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: University of Chicago Press.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Williams, D. S. (1972). Computer program organization induced from problem examples. In H. A. Simon & L. Siklossy (Eds.), *Representation and meaning: Experiments with information processing systems* (pp. 143–205). Englewood Cliffs, NJ: Prentice-Hall.

Appendix

Classification of Raven Problems by Rule Type

			% error				Lesioned model			
								Working memory limit by no. rules		
Raven no.	Taxonomy of rules in a row	No. rule tokens	Exp. 1a (n = 12)	Exp. 1b (n = 22)	FAIRAVEN	BETTERAVEN	No D2Val rule	3	4	5
I-2	Pairwise, constant	2	0	N/A	+	+	+	+	+	+
I-6	Pairwise, constant	2	8	N/A	+	+	+	+	+	+
I-7	Distribution of 3, constant	2	42	14	+	+	+	−	−	+
I-8	Distribution of 3	2	17	18	+	+	+	+	+	+
I-9	Distribution of 3, constant	3	42	5	+	+	+	+	+	+
I-10	Addition, constant	2	25	14	+	+	+	+	+	+
I-12	Subtraction	1	42	9	+	+	+	+	+	+
II-1	Distribution of 3, constant	3	8	9	+	+	+	−	−	+
II-3	Pairwise, constant	2	0	9	+	+	+	−	−	+
II-4	Pairwise, constant	2	8	5	+	+	+	+	+	+
II-5	Pairwise, constant	2	8	0	+	+	+	+	+	+
II-6	Pairwise, constant	2	0	0	+	+	+	+	+	+
II-7	Addition	1	17	14	+	+	+	−	−	+
II-8	Distribution of 3	2	17	18	+	+	+	−	+	+
II-9	Addition, constant	2	0	9	+	+	+	+	+	+
II-10	Pairwise, constant	2	17	5	+	+	+	+	+	+
II-12	Subtraction	1	0	9	+	+	+	+	+	+
II-13	Distribution of 3, constant	3	50	32	+	+	+	−	−	−
II-14	Pairwise, constant	2	8	9	+	+	+	+	+	+
II-16	Subtraction	1	17	41	+	+	+	+	+	+
II-17	Distribution of 3, constant	2	17	23	+	+	+	+	+	+
II-18	Unclassified[a]	N/A	42	N/A	−	−	−	−	−	−
II-19	Unclassified[a]	N/A	33	N/A	−	−	−	−	−	−
II-22	Distribution of 2	3	42	45	−	+	−	+	+	+
II-23	Distribution of 2	4	33	32	−	+	−	−	+	+
II-26	Pairwise, distribution of 3	2	50	67	−	+	−	+	+	+
II-27	Distribution of 3	2	42	36	+	+	+	+	+	+
II-29	Distribution of 3	3	75	95	−	+	−	+	+	+
II-31	Distribution of 3 & 2	4	42	55	−	+	−	−	+	+
II-32	Distribution of 3 & 2	4	75	73	−	+	−	+	+	+
II-33	Addition, subtraction	2	50	63	−	+	−	−	−	−
II-34	Distribution of 3, constant	4	58	73	+	+	+	+	+	+
II-35	Distribution of 2, constant	4	67	27	−	+	−	−	−	−
II-36	Distribution of 2, constant	5	83	N/A	−	+	−	−	−	−

Note. D2Val = distribution of two values; N/A = not applicable; plus signs indicate correct solution; minus signs indicate incorrect solution.

[a] Problem was not classifiable by our taxonomy.

Received September 18, 1989
Revision received February 26, 1990
Accepted February 26, 1990 ■