RUNNING HEAD: COMPLEX INFORMATION PROCESSING

Complex Information Processing and Intelligence

David Lohman

University of Iowa

Draft of

Wednesday, November 20, 2002

Outline

Complex Information Processing and Intelligence

(INTRODUCTION)...pp. 5-14

METHODOLOGY...pp. 14-28

PROCESS MODELS OF REASONING TASKS...pp. 28-54

Complex Information Processing and Intelligence

Theories of human intelligence must explain those complex human behaviors that are most commonly understood as its indicants. Thus, the central facts to be explained by a theory of intelligence must go beyond faster or more efficient processing of elementary tasks, for example, or the efficiency of biological processes and inherited structures, or the influence of individuals, environments, or even cultures. Rather, a theory of intelligence must explain the writing of novels, the solving of complex mathematical problems, the designing of skyscrapers and microchips, and the myriad other forms of complex cognition valued by society. In short, an understanding of how individuals solve complex tasks, and an explanation of why they differ so markedly in their ability to do so, is in fact central to any theory of intelligence.

Cognitive Tests as Cognitive Tasks

But where to begin? There are many thousands of complex tasks, each of which might be considered an indicant of intelligence (especially by those who excel in accomplishing the task!). Correlational studies of human abilities offer a reasonable starting place, since they have identified groups of tasks that consistently measure abilities viewed as indicants of intelligence--both by psychologists and lay persons. Estes (1974) argued that we start by examining intelligence tests:

> Rather than looking to learning or physiological theory or some correlate of intelligence, I should like to focus attention on intellectual activity itself. By bringing the concepts and methods of other disciplines to bear on the analysis of intellectual behavior we may come to understand how the conditions responsible for the development of its constituent processes and the manner of their organization lead to variations in effectiveness of intellectual functioning. If this approach has appeal in principle, we need next to consider just what behaviors to analyze in order to be sure that the activity we are dealing with is closely related to that involved in the measurement of intelligence. The simplest and most direct approach, it seems, is to begin with the specific behaviors involved in responding to items on intelligence tests. (pp. 742-743)

The vast majority of research on individual differences in intelligence as it relates to behavior on complex tasks has followed Estes' suggestion, even though some have questioned the wisdom of this decision. In this chapter, then, I will examine research on test-like tasks modeled after item-types commonly used on intelligence tests. I focus especially on measures of reasoning, particularly inductive reasoning, in part because reasoning tests have been studied extensively and in part because inductive reasoning is the primary ability most commonly associated with G. Gustafsson (19) claims, for example, that General Mental Ability (G) can be equated with General Fluidity Ability (Gf), which in turn can be equated with Inductive Reasoning (I). Sternberg (1986) makes a similar point:

> An interesting finding that emerges from the literature attempting to relate cognitive task performance to psychometrically measured intelligence is that the correlations of task performance and IQ seems to be a direct function of the amount of reasoning involved in a given task, independent of the paradigm or label given to the paradigm.... Thus, reasoning ability appears to be central to intelligence. (pp. 309-310)

Even those who believe that there is more to G than inductive reasoning would agree that reasoning is a crucial aspect of any understanding of human intelligence.

Although many different tasks have been used to measure inductive reasoning, a few are used much more commonly than others: analogies, matrix problems, series completions, and classification. Some reasoning batteries also contain tests that also measure verbal reasoning through sentence completion tests, sentence comprehension tests, and even vocabulary. Others include more specific spatial tasks, such as form boards or paper-folding tests. And others use quantitative tests that require examinees to make relational judgments (such as greater than or less than) between quantitative concepts, or to determine how numbers and mathematical operators can be combined to generate a product. Examples of these different item types are shown in Figure 1.

Insert Figure 1 about here

In addition to the identification of clusters of tasks that define particular ability constructs, correlational studies show how these tests and the factors they define are related to one another. There is now broad consensus that these relations can be represented hierarchically (Carroll, 1993; Gustafsson, 1988). Even more suggestive for the present discussion, however, was the demonstration that hierarchical factor models are conformable with a radex model. The radex is produced by treating test intercorrelations as distances, which are then scaled in two or three dimensions using nonmetric, multidimensional scaling. The resultant scalings show three important features, all of which are illustrated in the idealized scaling shown in Figure 2 (from Snow, Kyllonen, & Marshalek, 1984). First, tests cluster by content, which typically appear as verbal, spatial, and symbolic/quantitative slices of a two-dimensional radex. Second, tests and test clusters that define broad factors tend to fall near the center of the radex. More specific primaries fall near the periphery. Indeed, in a well-balanced battery of tests, tests that define G fall near the center of the plot. Third, task complexity is roughly related to distance from the center (or G). This suggests that one key to a theory of G, then, may be an understanding of the complexity gradients that emanate from G to more peripheral or specific abilities.

Insert Figure 2 about here

In short, correlational studies human abilities have guided investigations of the processes that generate intelligent behavior through (a) the identification of ability constructs that are broadly predictive of performance in non-test situations, (b) the isolation of tests that consistently measure these abilities and thus may constitute interesting cognitive tasks, and (c) the display of inter-construct and intertask relationships (such as the centrality of G and the apparent complexity gradients) that need to be explained by a theory of intelligence.

Complexity

But what is complexity? How might it be defined and measured? In the history of research on human intelligence, the "simple" reaction time and perceptual-motor tasks used by Galton and J. Mc. Cattell were distinguished from the more complex tasks used by Binet. "The early mental tasks were predominantly sensori-motor or very simple in nature.... [Then, as now,]

complex mental processes were believed to be best understood by analyzing them into their elementary components, usually of a sensory nature" (Anastasi & Foley, 1949, pp. 14-15). However, Binet and Henri (1896) argued that more complex tests were needed to measure intelligence. The battery they proposed included tests of memory, mental imagery, imagination, attention, comprehension, suggestibility, aesthetic appreciation, moral feelings, muscular force and force of will, and motor ability and visual discrimination. The success of Binet's test and the seeming failure of the Galton-Cattell approach (Sharp, 1898-1899) established a working definition of task complexity for differential psychologists for the next 70 years: complex tasks were those that required comprehension, judgment, reasoning, like Binet's tests; simple tasks were those that measured basic sensory and motoric processes. It was not until the advent of modern cognitive psychology--particularly the information-processing paradigm--the more precise definitions were offered.

Information Processing

"Information processing" does not label a unified approach, but rather a spectrum of researchers and theorists who use a variety of methods to study an equally diverse array of problems. On this view, cognitive tasks, including psychological tests, can be analyzed to reach deeper understanding of the mental processes and content knowledge that comprise complex performance. All information-processing models of thought posit one or more sequences of processing steps or stages in which cognitive operations are performed on incoming or stored information. The once fuzzy notion of "cognitive process" is thus concretely operationalized as a particular cognitive transformation performed on a particular mental representation. Some information-processing models are simple constructions with only one or two parameters to reflect the functioning of different processing stages in performance on paradigmatic laboratory tasks. Others are mathematical models of more complex tasks that appear important in their own right outside the laboratory. Some take the form of computer programs for complex tasks that reach a level of detail far more explicit than most mathematical models (Anderson, 1983, 1993; Newell & Simon, 1972). This detail can be overwhelming, however, unless one has some way

of discovering generalizations that hold across time, task, and persons. In other words, suppose one starts with a full accounting of how a particular person solves a particular task on one occasion. Some of these processes may not occur if the same task were administered at another time (even assuming no specific memory of the first solution). More importantly, even those processes that are common across occasions may not generalize to other tasks or to other persons. Indeed, the subset of processes that produce individual differences that generalize over time and tasks is likely to be but a tiny fraction of the processes required to specify the behavior of a single individual on one task on one occasion. It is important, therefore, to find a level of detail that explicates but does not overwhelm, that can capture useful generalizations that hold across families of tasks, and above all, that can represent individual differences in success in solving problems on these tasks.

General Features of Information Processing Models

There is a tradeoff between process and structure in all cognitive models. Models that specify many structures require fewer processes, and conversely. Of the many structural features of information-processing models that could be discussed, three seem particularly important for any discussion of relations between complex information processing and intelligence. First is the distinction between different memory systems--particularly working memory and long term memory. Second is the distinction among different types of knowledge representations, particularly between declarative knowledge and procedural knowledge, but also among different types declarative knowledge codes. And third is the distinction between controlled and automatic processing.

Memory systems. The earliest information-processing models posited one or more sensory buffers, a limited-capacity short-term memory, and an unlimited long-term memory. However, the old notion of a relatively passive, limited-capacity short-term memory has been replaced by a more active working-memory system that not only holds and manipulates information (Daneman & Carpenter, 1980), but also attends selectively to one stimulus while inhibiting another, coordinates performance on tasks, and switches strategies (Baddeley, 1996).

In fact, the characteristics

Baddeley now posits for supervisory attentional system in working memory look much like the sort of executive processes Snow (1978) and Sternberg (1977) hypothesized as essential features of intelligence.

Oberauer et al. (1996) summarize these processes as (1) simultaneous storage and processing, (2) supervision or monitoring, and (3) coordination. Although there is substantial agreement on the first function (simultaneous storage and processing), there is disagreement whether both storage and processing share common resources (Just & Carpenter, 1992), or draw on separate resources (Halford, Maybery, O'Hare, and Grant, 1994). Supervisory functions in most models include not only a monitoring function, but also an inhibition function (which was also the central component in Thurstone's [1924] earliest theory of intelligence). Finally, at least three aspects of coordination have been proposed: (a) information coordination across sensory modalities (Law, Morrin, & Pellegrino, 1995); (b) coordination of successive mental operations into a sequence (Hagendorf & Sá, 1995); and (c) coordination of elements into a coherent structure or mental model (Johnson-Laird, 1983; Oberauer, 1993).

Baddeley (1986) also claims that different memory systems are employed for verbal and spatial tasks. Some have also attempted to separate numerical from verbal content, in keeping with the long-established distinction among verbal, numerical/symbolic, and spatial abilities.

Declarative and procedural knowledge. A second important distinction is between declarative and procedural knowledge, or more simply, between content knowledge and process knowledge (Anderson, 1976, 1983, 1993; Greeno, 1978; Ryle, 1949). "Declarative knowledge" is factual knowledge about the meaning or perceptual characteristics of things, from anecdotal memories about daily events to the highly organized conceptual knowledge of some subject-matter. A novice's knowledge of a newly-learned theorem of geometry, for example, is declarative knowledge. Anderson (1983, 1993) also distinguishes between two types of declarative knowledge representations: a meaning-based memory code (abstract propositions) and a variety of perception-based codes (linear orders, image, etc.). The dominance of the meaning-based code

parallels the ubiquitous G factor in individual differences research, whereas the specialized codes correspond to at least some of the major primary ability factors, particularly verbal fluency and spatial ability.

"Procedural knowledge" is knowledge of how to do something, from pronouncing a word or driving a car, to transforming and rehearsing information in working memory, to assembling new methods for solving problems and monitoring the effectiveness with which these methods are implemented. Organizing words into a sequence to express a particular idea, for example, requires procedural knowledge, as does the mental rotation of a figure into a new orientation.

All cognitive tasks require both declarative and procedural knowledge, and in the abstract the two can be seen as one (Greeno, 1978; Winograd, 1972, 1975). Procedural knowledge can often be stated declaratively and declarative knowledge can become proceduralized with practice. Nevertheless, tasks differ in the demands they place on one or the other type of knowledge. For example, tests that sample factual knowledge in a domain place heavy demands on the examinee's store of declarative knowledge, its organization, and its retrievability. On the other hand, tests of inductive reasoning or spatial visualization ability place heavier demands on the examinee's execution of certain cognitive procedures or skills. Some tasks require complex mixtures of factual knowledge and cognitive skill (e.g., mathematical reasoning tests).

Controlled and automatic processing. The controlled versus automatic processing distinction refers to the degree to which a knowledgeable or skillful performance requires conscious attentional resources for successful operation. Automatization occurs when cognitive tasks are consistent in their information-processing demands such that a transition from controlled to automatic processing can occur with practice. Using arithmetic facts is a controlled process for a child. For the practiced adolescent, their use has become automatic. Automatization is thought to free attentional resources for other controlled processing. Tasks with inconsistent information-processing requirements remain consciously controlled; automatization is not possible. Tests of ability and achievement may emphasize either type of processing, or some mixture, and a given test may vary on this continuum as a function of where on a learning curve it is administered

(Ackerman, 1986). A perceptual speed test, such as "Number Comparison" may reflect automatic processing for most adults, but it may reflect controlled processing among children. A test composed of algebra word problems will likely require both kinds of processing. One important characteristic of reasoning problems is that they require controlled processing (Sternberg, 1986).

Information Processing and Task Complexity

The information-processing approach permits much greater specificity in hypotheses about complexity . Nevertheless, even within this paradigm, there is no simple way to distinguish complex information processing from simple information processing. For example, those who study reaction times properly speak of tasks that require a discrimination ("respond only when the red light is lit") as more complex than tasks that do not ("respond when any light is lit"). Clearly, "complexity" is a relative term. Simon (1979) argues that information-processing psychology has been pursued at two levels: at the level of the “immediate processor” and at the level of “relatively complex human performances” (p. xi). Investigations of memory-scanning or sentence-picture verification (see, e.g., Clark and Chase, 1972) characterize the former approach, whereas the Simon and Kotovsky (1963) studies of letter series tasks are typical of the latter. The duration of primitive processes are typically tens or hundreds of milliseconds in studies of the immediate processor, and of seconds--even tens of seconds--in studies of more complex behavior. Thus, although there is more at stake here than the nature of the dependent measure, classification of investigations by the nature of the dependent measure used affords a useful way to parse the domain.

Dependent measures. At the broadest level, one way to distinguish among levels of complexity is to examine the type of dependent measure used to index performance. Figure 3 shows such a continuum. At the simplest level, response latency is most useful. Errors are infrequent and may reflect nuisance factors (such as speed-accuracy tradeoff or inattention). Next come eye fixations. Again, fixation patterns are generally most informative when tasks are of relatively short duration, especially when comparing the performances of many subjects. However, performance need not be error-free. In fact, eye fixation patterns that differentiate

Insert Figure 3 about here

between correct and error trials may suggest how errors occurred. Next come protocol analyses. Subjects are asked to think aloud while solving a task. Verbal utterances are then analyzed to infer processes. Next comes a classification or analysis of the response itself, typically across many items or trials. On forced-choice tests, the classification scheme is often simply "correct" or "incorrect," and sometimes "attempted." However, open-ended responses can be classified in many ways. For example, responses on many of the tasks studied by developmental psychologists (such as balance-beam problems) can be classified according to the rule or rules seemingly used to generate the response. At a more complex level, essays can be scored for organizational structure or other emergent features (e.g., Biggs & Collis, 1981).

Sources of difficulty. In order to understand why subjects differ on a task, one must first understand what makes items difficult. On the typical reasoning test, easiest items will be answered correctly by 90 percent (or more) or the examinees, whereas the most difficult items may be answered correctly by only 30 percent of the examinees. Items vary not only in difficulty, but also in the particulars of what makes them difficult. By design, a good psychometric test may be a veritable hodgepodge of different sources and levels of difficulty. The first task, then, is to tease these sources apart--typically by designing item pools in which sources of difficulty vary systematically (and orthogonally, if possible). The advent of IRT models (see chapter __) in which items from different forms of a test may be placed on a common difficulty scale provides other options. Suppose a particular source of difficulty is reflected in only one or two items on one form of the verbal analogies subtest of the SAT. Placing many forms of the test on the same scale allows one to estimate the effects of that source of difficulty from a much broader sample of items. This is particularly useful on verbal items, given the large number of ways in which items can differ.

Sheehan (1997) studied sources of difficulty on the SAT Verbal Reasoning Test in this way. The dependent measure was an IRT estimate of item difficulty and the independent variables, various item characteristics hypothesized to influence difficulty. Sheehan (1997) used

a tree-based

regression procedure to assign items to nonoverlapping clusters. Figure 4 shows one set of results for the reading comprehension items. Clusters at the highest level (vocabulary in context, main idea, inference, and application/extrapolation) were specified in advance. Clustering at lower levels was determined by the regression procedure. Some clusters are composed of items of uniform difficulty, whereas items in other clusters vary substantially in difficulty. Not captured

in this sort of analysis, of course, is information about the extent to which different subjects or groups of subjects would be equally well characterized by the same clustering solution. Nonetheless, the analysis goes considerably beyond global statements about an undifferentiated item difficulty. Just because items can be nicely ordered on a single scale does not mean that they are psychologically homogeneous.

Insert Figure 4 about here

Methodology

Introspection, Retrospection, and Think-Aloud

Inferences about mental processes require a rich observation base. The first task of the cognitive scientist is thus to increase the density of observations between stimulus and response. The simplest and oldest method for doing this is to ask participants to introspect on their processing while attempting a problem and, after making a response, to summarize retrospectively their observations of self. However, the retrospective report will usually be an edited account that omits unproductive avenues of thought, or simply orders events in a more logical manner. In spite of these limitations, retrospective reports have long contributed usefully to inferences about processes (see, e.g., Bloom & Broder, 1950). A more common procedure is to ask the participant to think aloud while solving the problem. The basic assumption is that subjects can do this without altering their problem-solving processes as long as the information that is reported is already in working memory and does not require additional processing (Ericsson and Simon, 1984). This is often a reasonable

assumption if subjects are verbally fluent, if the task is one that demands or affords verbal labeling of objects and transformations, and if subjects are well practiced. However, some subjects find a think-aloud protocol intrusive and difficult. Francis Galton spoke for many of these people when he wrote:

> It is...a serious drawback to me in writing, and still more in explaining myself, that I do not so easily think in words as otherwise. It often happens that after being hard at work, and having arrived at results that are perfectly clear and satisfactory to myself, when I try to express them in language I feel I must begin by putting myself on another intellectual plane. I have to translate my thoughts into a language that does not run very evenly with them. (in West, 1991, p. 179)

In spite of such limitations, introspective, retrospective, and think-aloud reports are such a rich and important source of information on cognition that they are used in studies of complex problem solving--either as a primary dependent measure (as, for example, in comparisons of expert versus novice problem solvers) or a secondary dependent measure (as, for example, in attempts to develop computer programs that simulate human problem solving).

Computer Simulation

One of the best methods for understanding how individuals solve a task is to attempt to simulate their behavior with a computer program. Early claims about the potential of the computational approach to cognition often relied on the success of a particular program in solving items like those found on ability tests. A key distinction in this work was between algorithms and heuristics. An algorithm specifies a series of operations that, if performed correctly, will produce a solution to a problem. For example, children are taught an algorithm that will allow them to solve long-division problems in arithmetic. A heuristic, on the other hand, is a rule of thumb that specifies a series of steps or operations. Unlike an algorithm, however, the heuristic may or may not lead to a correct solution of the problem. Many of the early attempts to simulate human problem solving sought to develop computer programs that would apply general problem solving

heuristics to a wide range of problems. For example, the General Problem Solver (Newell & Simon, 1961) implemented a means-ends heuristic in which the problem solver determines the current state, the goal state, and selects a path or operation that will effect the greatest reduction in distance between the two. Means-ends analysis, then, is a working-backwards heuristic, since the problem solver starts with the goal state and tries to work backwards to the present state.

Studies of expertise, however, demonstrated that although novices tended to use working-backward heuristics such as means-ends analyses, experts tended to use working-forward strategies (Simon & Simon, 1978). For example, experts in algebra may begin by rearranging terms without any clear understanding for how such a transformation gets them closer to the goal. However, once the terms are rearranged, another transformation will generally be suggested by the new pattern and the process repeated until the problem is solved. Such behavior is better characterized by a series of production rules that are applied automatically and consecutively. It suggests that domain-specific knowledge enters more prominently into the problem-solving behavior of experts than of novices.

Many of the early simulation programs ignored or at least downplayed the issue of knowledge. However, studies of human experts consistently showed that they relied extensively on stored problem patterns, not only in solving artificial problems (such as chess) but also in more natural domains such as language processing (Schank, 1980). The pendulum thus swung from the knowledge-lean programs of Newell and Simon (1961) to knowledge-rich expert systems. Indeed, even the architecture of different expert systems seemed to vary by the type of problem the system was expected to solve (Sternberg, 1990). Nonetheless, there are common structures and processes that differentiate between more- and less-intelligent systems. Sternberg (1990) identifies seven: multiple indexing, higher-order knowledge structures, beliefs, drawing inferences, levels of understanding, and learning mechanisms. Multiple indexing refers to the need to tag or index new information in many different ways (or simultaneously to store it in many different “places”). Multiple indexing is important because later recall and use of that information depends critically on how elaborately it was processed on input. Higher-order knowledge structures refer to scripts

or schemes. Schemes are crucial in allowing the system to organize incoming information, to determine which information is central and which is peripheral, and to go beyond the information given

in making inferences about the situation. Beliefs are typically implemented as very high-level knowledge structures and thus function like schemes but at a more general level in the system. For human problem solvers, beliefs are often imbued with affect as well (Damasio, 1994).

More intelligent systems are also better able to make good inferences. One way to accomplish this is through high-level knowledge structures such as schemes: unfilled slots in the scheme are filled with default values. Another is through control mechanisms that index the extent to which new information deviates from the normal or expected. Schank (1978) argues that an intelligent system will make good use of such "interesting" information. Schank (1984) also argues that levels of intelligence are related to levels of understanding, from simply making sense of a situation through levels in which systems justify their actions in terms of goals, alternative hypotheses, and knowledge used, to a level of "complete empathy" in which the system's explanation would include an account of its feelings and emotions.

Systems also vary enormously in their ability to learn. Exactly how this occurs varies across systems. At the simplest level, systems learn by adding new knowledge to an existing knowledge base, modifying the strength of association among elements in that base, and reorganizing the knowledge base. Other systems allow for learning of new skills and heuristics, and for the conversion of fact knowledge to skill knowledge.

Componential Analysis

A detailed information-processing model is the single most important outcome of a cognitive analysis or computer simulation of a task. In the best methods for reaching this goal, the investigator specifies the details of each component process or stage, and how these component processes combine, then finds ways to operationalize the model, then evaluates the model by comparing how well it accounts for variation in the data with that of rival models, then revises

and reevaluates accordingly. This is an important departure from methods in which an investigator attempts to infer a full model from one or two statistically significant effects (see Greeno, 1980a, for further discussion).

One of the earliest and most influential methods for accomplishing this sort of task decomposition and model testing was formulated by Sternberg (1977) in a method he called componential analysis. Early experiments examined analogies; in later experiments, the method was applied to a wide range of reasoning tasks. Sternberg proposed that analogy items--such as "doctor:patient::lawyer:(a) client, (b) judge"--required at least five component processes: encoding the various terms (here referred by the letters A:B::C:D_1,D_2), inferring the relationship between A and B, mapping the relationship between A and C, applying the A-B relationship to C to generate an answer, comparing the constructed answer with the alternative D answers provided, and responding. By the definition of analogy, however, the nature of the A to B transformation will

be the same as that of the C to D transformation. For example, in a verbal analogy, if A is the opposite of B, then C must be the opposite of D. Thus, independent variables representing the type of inference or amount of transformation required in the inference stage would be perfectly correlated with corresponding independent variables for the application stage, so only one set of variables could be entered into the regression equation. This is a problem if one wants to test the assertion that two component processes are required, or to estimate the speed or efficiency with which these separate component processes are executed for individual subjects. Perhaps, some subjects quickly infer the correct relationship between A and B but have difficulty remembering the relationship and thus have difficulty applying it to C in order to generate an ideal answer D.

Several experimental manipulations have been proposed to unconfound these correlated variables (Sternberg, 1985). One method is called precuing. Here, the subject first examines a part of the problem for as long as necessary, then signals readiness to see the remainder of the item. Two latencies are obtained: time taken to view the precue, called a cue latency (which is

sometimes discarded), and time taken to solve the remainder of the problem, called a solution latency. For example, given the analogy

A:B::C:D (true or false?),

one could precue with

A:B:: (measure T_1)

and then present

C:D (true or false) (measure T_2 and subject's response).

The precuing time (T_1) would include time for encoding A and B, and for inferring the relationship between A and B. The separate solution time (T_2) would include the time for encoding C and D, and for the mapping, application, and comparison-response processes. Precuing only the first term (A:) or the first three terms (A:B::C:) would permit separation of other component processes.

Precuing has probably been less frequently used than experimentally decomposing a complex task into two or more consecutively presented steps. Response latencies (and sometimes, response errors) are estimated for each step or subtask. Models are then fitted to each step (e.g., Kyllonen, Lohman, & Woltz, 1984) or to all steps simultaneously (e.g., Sternberg, 1977). When the latter procedure is used, model fits may be inflated because any variable that distinguishes among cue conditions will account for variance in latencies to different items. For details on the modeling procedures, see Sternberg (1977) and Pellegrino and Lyon (1979).

Strategies and Strategy Shifting

The first results from componential analyses were process models for several kinds of reasoning tasks. But the assumption that the same model should fit all subjects on a given task was quickly recognized as a useful first approximation that might be unwarranted on some tasks. Different process models might thus be needed to fit the performance of different individuals. Both Sternberg (1977) and Snow (1978) had predicted that strategic variations would be an important source of individual differences in complex information processing. Although such differences in strategy are common on complex tasks, they occur on simpler tasks as well. For

example, Cooper (1982) distinguished different strategies on a spatial comparison task. She also discovered that individuals sometimes could change strategies if induced to do so.

MacLeod, Hunt, and Mathews (1978) demonstrated a similar divergence of strategies on a sentence verification task. The performance of some subjects was well fit by a linguistic processing model, whereas the performance of other subjects was better described by a spatial processing model. In other words, a linguistic processor might read the presented sentence, encode it in memory linguistically, and also the picture when it appeared, and then compare these linguistic descriptions; a spatial processor might use the sentence to visualize the expected picture and then compare this image to the picture when it appeared. Verbal and spatial ability test profiles for the

two groups differ in expected ways, suggesting that strategy choice was systematic. However, it was also clear that a subject could change from one strategy to the other if asked to do so.

Sternberg and Weil (1980) also conducted a training experiment in which linguistic versus spatial strategies were contrasted for reasoning in linear syllogisms. Linear syllogisms are problems such as "Mary is taller than Sue. Sue is taller than Amy. Who is shortest? Mary, Sue, Amy." Componential models were fit to identify four different strategy groups. Correlations with reference ability tests showed strikingly different patterns across groups. Success with the linguistic strategy correlated with verbal but not spatial ability, the opposite pattern occurred with the spatial strategy, and a mixed strategy group showed correlation with both abilities. Those using a fourth strategy, a simplified algorithmic procedure, showed reduced correlation with ability.

Kyllonen, Lohman, and Woltz (1984) next showed how componential models can be generalized to account for cases in which subjects not only use different strategies but shift between them during task performance. Although many investigators had argued that subjects seemed to shift strategies as items became more difficult (Mulholland, Pellegrino, & Glaser, 1980; Bethell-Fox, Lohman, & Snow, 1984), no one had yet tested this within the framework of componential analysis. Kyllonen et al. (1984) identified three kinds of ability-strategy relationships: In a Case I

relationship, ability limits strategy selection. In Case II, strategy choice is unrelated to ability, but the effectiveness of implementing a strategy depends on ability. In Case III, ability both limits strategy and predicts performance within strategy groups. Evidence for all three cases was found in componential analyses of a complex form board task that contained three steps. In the first step, subjects were required to memorize a geometric figure. In the second step, they were required to combine this first figure, which was no longer in view, with one or two new shapes displayed on a screen. In the third step, they were shown another figure and asked whether the two or three previously shown shapes would combine (in the order indicated) to form this final shape.

Models for each of the three steps (called encoding, synthesis, and comparison) were constructed from retrospective reports from experimental subjects, introspections of the experimenters, and the literature on spatial cognition. Each was then tested by regressing latencies for each step on independent variables that estimated the amount or difficulty of each hypothesized process. Independent variables were formed by coding objective features of items or by obtaining ratings of the desired characteristic from other subjects. Two types of models were tested: single-strategy models and strategy-shift models. Single-strategy models presume that the subject solved all items in basically the same way. Strategy-shift models, however, presume that the subject uses different strategies to solve different types of items.

The results of fitting these sorts of models to response latencies for each of the three task steps showed two important effects. First, for each task step, different subjects solved different items in predictably different ways. Second, solution strategies were systematically related to the profile of scores on reference ability tests. This is shown in Figure 5. For example, for the synthesis step, subjects who were best fit by the most complex strategy-shifting model had the highest average ability profile. In other words, high-G subjects showed the most flexibility in making within task adaptations in their solution strategies. Subjects who followed more restricted strategies had more extreme profiles, either always synthesizing figures (high on spatial but low on verbal ability) or synthesizing only those figures presently in view (very low spatial, but average on verbal ability).

Insert Figure 5 about here

Other studies have supported the utility of strategy-shift models and also have shown that seemingly minor variations in task demands can have a pronounced impact on how subjects solve items on spatial and figural reasoning tasks (Embretson, 1986; Lohman & Kyllonen, 1983; Lohman, 1988). Bethell-Fox et al. (1984) also demonstrated that strategy shifting within persons in geometric analogy tasks could account for differences between previous models offered by Sternberg (1977) and Mulholland, Pellegrino, and Glaser (1980). Furthermore, analyses of eye movements during performance of analogy items added evidence that subtle individual differences in strategy shifting, not previously incorporated in componential models of reasoning about geometric forms. Although spatial tasks seem particularly susceptible to alternative solution strategies, verbal tasks are not exempt. Marquer and Pereira (1987) have reported that substantial strategy shifting occurs even on the sentence verification task used previously to contrast strategies that were assumed to be stable. For example, more than two-thirds of the subjects in their study exhibited nonrandom shifts during performance. It is possible then that strategy shifting is a key aspect of individual differences in verbal as well as spatial and reasoning task performance (see, e.g., Spiro & Myers, 1984).

Strategy shifts can represent noise or substance, depending on the goal of test analysis. They contribute noise if the goal is to estimate a particular set of process parameters (e.g., rate of rotation, speed of lexical access) from a sample of items. They represent substance important for theory if, at some point, a higher level of ability means having a more flexible approach to problem solving. In either case, in interpreting research on abilities, it is incorrect to assume that all items or problems in a task are solved in the same way and that subjects differ only in the speed or power with which they execute a common set of processes. It is also incorrect to assume that subjects may be typed by strategy, or even that they shift strategies in the same way. At least, general abilities such as verbal comprehension, reasoning and spatial visualization--appear to be more complex than this. Higher levels of performance in these ability domains seems to involve at least some flexible strategy shifting. On the other hand, it is abundantly

clear that ability is much more than strategy. Indeed, the most able subjects often show little evidence of strategy shifting, most probably because they have no need to do so on the problems presented. Thus, although strategies and strategy shifting are important, it is unclear how much of the variation in G can be attributed to such processes and how much it reflects other factors such as differences in attentional resources or knowledge, for example.

Modeling Response Errors

Regression models. Problems on ability tests tend to be hard. Individual differences are reflected in how many problems examinees answer correctly, not how rapidly they finish. For example, time limits on the Cognitive Abilities test (Thorndike & Hagen, 1993) are such that 98-99% of students respond to all items each of the nine subtests. Indeed, complex tests that load highly on G tend to be less speeded than tests that define more specific factors (or fall near the periphery of the radex). This is especially the case for spatial tests (Lohman, 1979). Thus, one of the earliest criticisms of the information-processing studies of abilities was that models that well described how rapidly subjects solved simple tasks might miss important aspects of individual differences in intelligence.

The regression procedures used to model response latencies assume that total response time is the simple sum of the time required to perform each of the hypothesized component processes. When modeling response accuracy, however, a probabilistic model (such as logistic regression) is more appropriate. An additive model makes little sense for response probabilities. A more plausible model would allow for conditional independence between components. In other words, the probability of executing component n+1 correctly is independent of the prior n components, but only if they were executed successfully (Embretson, 1995).

Goldman and Pellegrino (1984) provide an excellent introduction to models of response accuracy. They begin with the derivation of a fairly simple model for an analogy in which E stimulus elements are subjected to T transformations. They assume that there is some probability α of misrepresenting individual transformations and a similar probability β of misrepresenting individual elements, then the general equation for predicting error rate on a given class of problems would be

$$\text{Probability (error)} = 1 - (1-\alpha)^{T} (1-\beta)^{E}$$

A more complete model separates different aspects of item processing, such as the processing of the item stem and the processing of item alternatives. For example, in Alderton, Goldman, & Pellegrino (1985), accuracy of stem processing was estimated by either

a) presenting the A and B terms of the analogy, and asking the subject to generate a relational rule. If the rule was judged to be one that would lead to accurate solution, then inference is scored correct. (component process: inference)

b) presenting the A, B, and C terms, and asking the subject to generate a response. Response was scored correct if it was the target response or a synonym. (component processes: inference, mapping, application)

For alternative processing, the stem and five alternatives were presented.

c) If subject generated an incorrect relationship or alternative when shown only the stem but then chose the correct alternative when shown the alternatives, then recognition was coded.

d) If the subject generated a correct relationship or completion term when shown the stem, but then chose an incorrect alternative when shown the full item, then distraction was coded.

These separate probabilities were then combined for each subject in a single equation that predicted overall probability of a correct response on forced-choice items as the combined probability of correctly processing the stem (inference, application) and correctly processing the alternatives (recognition, guessing, distraction). The possibility that different subjects use different strategies is directly reflected in the estimated probabilities for each component in the model.

Error models can also be formulated to test for within-subject as well as between-subject shifts in solution strategy. For example, Embretson (1985, 1986; see also Embretson, Schneider, & Roth, 1986) used multicomponent latent-trait models to test the hypothesis that subjects attempted to solve items in more than one way. Embretson proposed that subjects first attempt to solve verbal analogy items using a rule-oriented strategy much like the strategy Sternberg hypothesized: they

infer a rule that relates the first and second terms of the analogy, apply this rule to the third term to

generate an answer, then compare this answer with the response alternatives. However, subjects switch to a secondary strategy if any part of the rule strategy fails. Three secondary strategies were proposed, based on associations, partial rules, and response elimination. In the association strategy, subjects choose the alternative that is most highly associated with the third term; this is sometimes also the keyed answer (Gentile, Kessler, & Gentile, 1969). In the partial-rule strategy, subjects infer only part of the rule that relates the first and second terms, but this partial rule may be sufficient to eliminate all distractors. In the response elimination strategy, subjects again infer only a partial rule, but this partial rule serves to eliminate only some of the distractors, so other item features or guessing must be used. Component scores for exact-rule construction, partial-rule construction, and response elimination were estimated by showing subjects the first two terms of each analogy and asking them to write the rule that described the A-to-B relationship. This procedure is the accuracy analog to the precuing procedure Sternberg used for response latencies. Exact rules stated the keyed relationships; partial rules stated only some of the keyed relationships, but served to eliminate all foils; response elimination was scored if the partial rule eliminated only some foils. A rule evaluation component was scored by providing subjects with the exact A-B rule, and asking them to select the correct alternative.

Different strategies were then formulated by combining these five component probabilities (plus another variable for overall probability of strategy execution) in different ways. For example, the probability that person j solves the full (T) item i by the rule strategy is given by

$$P_{rule}(X_{ijT}) = P_a P_{ij1} P_{ik2} \text{ where}$$

P_{ij1} is the probability of correctly inferring the exact A-B rule,

P_{ik2} is the probability of correctly applying the exact A-B rule,

P_a is the probability in the sample of applying the rule strategy.

The multiplicative relationship between component probabilities indicates that component operations are sequentially dependent: a failure on one component leads to a failure of the strategy. Also note that P_a does not vary over persons or items in this model.

Multiple strategy models were formed by summing the probabilities for individual strategies. For example, the probability of solving an item by first attempting but failing the rule strategy, and then selecting the correct response by choosing the alternative most highly associated with the C term is

$$P_{assoc + rule}(X_{ijT}) = P_{rule}(X_{ijT}) + P_{assoc}(X_{ijT})$$

where

$P_{assoc}(X_{ijT}) = P_c P_{ij3} (1-P_{ij1}P_{ij2})$,

P_{ij3} is the probability of choosing the correct answer by simple association,

$(1-P_{ij1}P_{ij2})$ is the probability that the rule strategy failed, and

P_c is the probability in the sample of applying the association strategy when available.

Adding probabilities for mutually exclusive strategies makes these models compensatory, like the strategy-shift models studied by Kyllonen, Lohman, and Woltz (1984) for response latencies.

Analyses showed that Model III (rule or partial rule strategy) and Model IV (rule strategy or response elimination) both predicted substantially more of the variation in estimated item difficulties than did Model I (rule strategy alone). However, all models were equally good predictors of subject differences. Thus, these models suggest that strategy shifts contributed significantly to item difficulty, but not to individual differences in overall performance in the task. This may reflect the fact that the probability of executing each strategy (P_a, P_b, or P_c) could not be separately estimated for each subject.

Rule space analyses. The rule space methodology was developed as a method for diagnosing errors (or "bugs") in component skills from observations of complex performances that require combinations of many skills (see Tatsuoka, 1995, 1997 for overviews). Initially the method was applied to computation skills in mathematics. Of late it has been applied to verbal tests as well, such

as the Critical Reading, Sentence Completion, and Verbal Analogies subtests of the GRE. Continued refinements and elaborations have resulted in complex set of procedures shown schematically in Figure 6.

Insert Figure 6 about here

The first step in conducting a rule space analysis is to identify the attributes hypothesized to be required by items on the test. In elementary mathematics, these would be component skills (such as "carry" in multicolumn addition). For more complex problems, attributes are often far less specific ("ability to synthesize information") and sometimes represent different aspects of a common skills (e.g., vocabulary skills are represented by the average word frequency of the five most difficult words and also by the average word length of words in a passage). The presence or absence of each attribute is then coded for each item on the test (this is the incidence matrix in Figure 6). All possible latent knowledge states are then generated from the incidence matrix using a procedure called a Boolean descriptive function. Each of these latent knowledge states corresponds to a particular pattern of scores on the test items. Observed score patterns are then mapped on to these ideal score patterns, with some allowance for "slips" or model misfit. Finally, for each subject classified in one of the knowledge states, the probability that the subject has mastered each attribute specified in the original list of attributes is computed.

In typical study, the procedure is applied iteratively--usually beginning with a long list of attributes that is reduced over iterations. Success of the model has been indexed by the percentage of test takers classified in one the latent knowledge states specified in the model (which typically number in the thousands), and by the regression of total score on the test on the set of attribute mastery scores. More convincing indices would be the probability that an individual is classified in a similar knowledge state on a parallel test, and the demonstrates that attribute mastery scores generalize across test forms and show convergent and discrimant correlations with external measures of the same or similar constructs.

This brief summary cannot begin to explain the rule space method. It is introduced here, however, as an illustration of a larger trend in research on individual differences in complex information processing. Limitations of the methods that were used to model performance on relatively simple test-like tasks have become more apparent over the years. More complex statistical methods have been developed in an effort better to model the enormous diversity of processing that occurs on complex tests such as the GRE, and also to take advantage of advances in measurement (such as IRT-based scales) that had been ignored in earlier efforts. There is a great paradox here, however, in that diversity one level is purchased at the price of uniformity at another level. For example, the basic assumption of the rule space procedure is that items can be unambiguously characterized as requiring or not requiring a particular skill. Thus, each column in the incidence matrix represents a strong statement about how all subjects must solve that item. Alternative strategies or compensatory mechanisms are not allowed. A person is expected to solve the item

only if he or she has mastered all of the component skills that are specified in the matrix. It may well be that such strong models can work when attributes are specified at a global level (e.g., "can synthesize information" or "can apply general background knowledge"). Such attributes are perhaps better seen as category labels for task demands (“can do X somehow”) rather than specific cognitive processes (“does X at Y rate”). Indeed, whether test takers can be classified into one or more knowledge states specified by the procedure may be less important for theory than the fact that collections of items used in these tests are being scrutinized for the demands they place on test takers.

Process Models of Reasoning Tasks

Hypotheses about Processing Complexity

Returning to the radex example in Figure 2, researchers working within the information-processing paradigm have offered several hypotheses to explain how tasks increase in complexity along these spokes: (a) an increase in the number of component processes involved in task performance; (b) an increase in the involvement of one or more particular components

(such as

inference); (c) an increase in demands on working memory or attentional resources; and (d) an increase in demands on adaptive functions, including executive or metacognitive controls (Snow, Kyllonen, & Marshalek, 1984).

Investigations of particular reasoning tasks sometimes address one of these hypotheses, but more commonly cut across them. Thus, although I will use these hypotheses to organize the discussion, the initial presentation of research is by necessity more task based. I begin with the early simulation of reasoning on series completion problems, then move to analogies, sentence completion, classification tasks, seriation tasks, and matrix tasks. I then summarize more general theories of reasoning that cut across task boundaries.

Series Completion

Series completion problems require the test taker to extrapolate the next member of a series of stimuli such as letters, numbers, or geometric figures. For example, one might be required to identify the next number in the series 1, 3, 6, 10, __ or the next letter in the series b, c, x, d, e, y, __.

Simon and Kotovsky (1963) proposed a computer simulation model for letter series problems of this sort. Their model contained two basic routines: a pattern generator and a sequence generation. The first corresponds to Spearman's "eduction of relations" and the second to his "eduction of correlates." The pattern generator was composed of subroutines for (a) detection

of the interletter relations; (b) identification of period length of the pattern, and (c) generation of a description or rule integrating these two aspects of the problem. Subsequent investigators often subdivided this last component into pattern description and extrapolation. In the Simon and Kotovsky (1963) theory, the rule or pattern description output by the pattern generator module served as input to sequence generator module, which applied the pattern description to the problem to generate elements that would come next in the problem. The knowledge base assumed by the program was limited to forward and backward knowledge of the alphabet, and the relational concepts of identity, next, and backwards-next (or reverse order).

Each of these aspects of problem description were shown to be related to problem difficulty. Identity relations were easier than next relations, which were easier than backwards-next relations. As in most problems that contain multiple sources of difficulty, however, the relative difficulty of different nonidentity relations depends on the location of the relation within the period. In particular, the difficulty of a next relation increases as it is further embedded in the pattern. In general, the longer the pattern description and the greater the demands presumably placed on working memory

in the identification of the rule, the more difficult the problem (Simon & Kotovsky, 1963; Holzman, Glaser, & Pellegrino, 1976). However, contrary to the model, period length does not appear to be related to solution accuracy.

Butterfield, Nielsen, Tangen, and Richardson (1985) revised and elaborated on the model to account for item difficulty in all possible series items, not for just the samples of items that might happen to appear on some particular test. They argue that a good theory how people solve series problems, and of what makes such problems easy or difficult to solve, must apply to representative samples of items from a defined universe that includes all known item attributes. As some attributes are over- or under-represented in the sample of items studied, the resulting theory of solution processes or item difficulty will be distorted and will vary unpredictably from study to study. Their theory discards period length as important in the representation stage, positing several levels of knowledge about what are called moving strings and predicting that item difficulty is determined by the string that is most difficult to represent. It also subdivides the memory load aspect of the continuation stage of performance. The theory accounts well for the data from the earlier work and from several new experiments.

LeFevre and Bisanz (1986) sought to provide a more detailed account of the processes involved in the "detection of relations" component and to determine the extent to which individual differences in tasks designed to measure this component predicted performance in the number series subtest of the Lorge-Thorndike Test of Intelligence (Lorge & Thorndike, 1957). LeFevre and Bisanz (1986) hypothesized that three procedures were used for detecting relations

among

numbers: recognition of memorized numerical series (e.g., 2 4 6 8), calculation (e.g., computing the interelement differences in a series such as 1 2 4 7 11), and checking (e.g., determining whether the last digit in "2 5 8 11 13" was encoded or calculated incorrectly, or instead marks the boundary between two periods). These hypotheses were investigated on very simple problems that combined procedures in different ways. Results showed that high-ability subjects used recognition of memorized sequences on a wider range of problems and calculated more efficiently than did low-ability subjects. High- and low-ability subjects did not differ on checking, although this may have been due to the low error rate (2.8% overall).

These studies of series completion tasks illustrate the range of analytic procedures used to infer process. The early studies of Simon and Kotovsky relied heavily on introspection and think-aloud protocols for theory generation. However, the test of the theory was the ability of the computer simulation to solve problems successfully and also to account for sources of difficulty in the task. The later work of Butterfield et al. (1984) and LeFevre and Bisanz (1986) focused on the analysis of response latencies and errors. Although simulation models continue to be an important source of evidence (e.g., Carpenter, Just, & Shell, 1990), methods for testing processing models against error and latency data have become much more popular. One reason has been the success of a set of methods for latency and error modeling known collectively as componential analysis, perhaps best illustrated in the research on analogical reasoning.

Models of Analogical Reasoning

Sternberg (1977) reported several investigations into the processes subjects use to solve analogies. Figure 7 shows flow charts for four alternative models of analogy performance he hypothesized. Component processes are identified inside the boxes; parameters reflecting the operation of each component are listed at the side of each box.

Insert Figure 7 about here

In Model I, the inference, mapping, and application components are all exhaustively applied, i.e., all attributes of the terms of the analogy are compared. In Model II, the application component is self-terminating, that is, after D is encoded, attributes are tested one at a time until a

correct attribute is found for response. In Model III, both mapping and application are self-terminating.

In Model IV, inference also becomes self-terminating. Model III best fitted the data in Sternberg's (1977) first experiments, accounting for 92, 86, and 80 percent of the variance in the latency for schematic-people, verbal, and geometric analogies data, respectively. A subsequent experiment with schematic-people analogies confirmed this order.

Geometric analogies. Mulholland et al. (1980) proposed a somewhat different model for geometric analogies. Borrowing from Evans (1968), they argued that subjects infer the relationship between A and B, infer the relationship between C and D, and then compose the two sets of relationships. Mulholland et al. (1980) refer to Sternberg's hypothesis as an infer-apply-test model and to their own as an infer-infer-compose model. Importantly, they studied true-false analogies, in which an infer-infer-compose strategy would make more sense than in the two alternative items Sternberg (1977) studied or the four alternative items Bethell-Fox et al. (1984) studied. Probably the most important result of their study, however, was the finding of a small but significant interaction between the number of elements or figures and the number of transformations (size change, rotation, doubling, etc.). Problems requiring multiple transformations on a single element were much harder than problems that required the same number of transformations on single elements. They argued that such items further burdened an already taxed working memory by requiring participants to retain in memory and then operate on the intermediate products of a transformation. (They also noted that the solution of "ambiguous" items [i.e., items in which the A-to-B transformation was nonobvious], although common on psychometric tests, were not included in their study.) They also speculated that strategic flexibility may be particularly important when attempting complex items.

The hypothesis that more complex items induce subjects to alter their strategies was addressed in a study by Bethell-Fox, Lohman, and Snow (1984). Bethell-Fox et al. (1984) presented both two-alternative items (as in Sternberg) and four-alternative items (as on some

mental tests). They also recorded eye fixations and administered an extensive battery of reference ability measures.

Several of their results were particularly noteworthy. First, unlike Sternberg (1977), they included a comparison component in all of their componential models. With the addition of this component, Sternberg's Models I through IV (see Figure 7) all fit the data equally well ($\underline{R}^2$ =.911 to .917 for two-alternative items, and $\underline{R}^2$ =.938 to .941 for four-alternative items). Mapping was a trivial component and was dropped, and two new components were added: spatial inference and spatial application. These components were only implicated on items involving spatial transformations. In other words, subjects appeared to activate and deactivate these processes across items.

Bethell-Fox et al. (1984) also followed up on Mulholland et al.'s (1980) suggestion to include items in which the nature of the A-to-B transformation was not immediately apparent (i.e., the most obviously correct answer did not appear among the response alternatives). The component Sternberg (1977) called justification accounted for approximately 25 percent of the variance in both latency and error data on these ambiguous items, but was not significant for nonambiguous items.

Analyses of eye fixations, however, showed that subjects often looked back to the A and B terms after inspecting C and the alternatives. Such lookbacks were particularly common on difficult, four-alternative items. This suggests that lookbacks were an integral part of solution strategy. Other results suggested that as problems increased in difficulty, subjects shifted from a strategy of constructing an ideal answer and comparing it to alternatives, to a more iterative systematic examination of stem and alternatives. Lower-ability subjects transited to this strategy sooner than their high-ability counterparts.

Verbal analogies. The declarative knowledge necessary to solve a geometric analogy is relatively easy to specify. Although the set of rules that transform one element into another is not small, it is not large either. Most items are constructed by applying one or more simple transformations (size change, rotation, shading change) to specific elements to produce specific

products. However, in verbal analogies, the transformation rules are often subtler and the elements far more variable. Further, variables hypothesized to index these relationships must often be obtained from ratings rather than from codings of objective features of stimuli.

Several classification schemes have been proposed for classifying verbal analogies by the type of semantic relations represented in the stem. These include class membership, function, location, conversion, part-whole, order-in-time, and property (Pellegrino & Glaser, 1982; Whitely, 1976). Location and function relations tend to be easier than others. Nevertheless, Pellegrino and Glaser (1982) argue that classification schemes do not predict difficulty, because they capture only the most salient relational feature. The ease or likelihood of identifying a relation varies enormously across items within each category. For example, in one study, a group of undergraduates generated responses to 150 stems (A:B::C:?). The probability associated with the most frequently generated response ranged from .10 to .90, indicating substantial variability in the typicality of generated answers, most of which were semantically appropriate (e.g., for "antler:deer::tusk," the most frequent response was "elephant," with "walrus" a distant second). As would be expected, items in which the constraints imposed by the stem restrict alternatives tend to be easier.

Building on the earlier work of Whitley (1976) and Chaffin and Herrmann (1984), Bejar, Chaffin, and Embretson (1991) proposed a more elaborate taxonomy of semantic relations. As shown in Table 1, the ten categories in their taxonomy are grouped into two higher-order categories of intensional and pragmatic relations. Intensional relations are said to be based solely on the meanings of the two words, and require an evaluation of the overlap or contrast of attributes of the two concepts. Pragmatic relations, in contrast, require knowledge of the world that goes beyond the meaning of the two words.

Insert Table 1 about here

In an analysis of 179 GRE verbal-analogy items, Bejar et al. (1991) found differences in the average difficulty of items in each of the ten categories. With one exception (Contrast Relations), the easier items were all classified as pragmatic relations. However, a follow-up study

by Diones, Bejar, and Chaffin (1996) failed to replicate this finding on SAT items. More importantly, mean difficulty in the earlier Bejar et al. (1991) study was inversely related to mean item discrimination
(r-biserial with GRE verbal composite). The authors show that this relationship between difficulty and discrimination can be eliminated by estimating discrimination using a coefficient that takes into account the differential attractiveness of different alternatives to examinees of different levels of ability. Put differently, one of the features of difficult analogy items is that they require careful discrimination among two (or more) plausible alternatives.

Buck, VanEssen, Tatsuoka, & Kostin (1998) investigated SAT verbal analogies using Tatsuoka's (1995) rule space methodology (see pp.___). The first step in this analysis requires the identification of attributes that are hypothesized to affect performance on the items. They identified a large number of variables, most of which indexed different aspects of vocabulary difficulty and semantic or conceptual complexity. These are shown in Table 2. The major determinants of semantic complexity were whether a word had multiple meanings and whether it referred to an abstract concept. Attributes of the relationship between concepts were the ability to recognize a negative rationale, a complex rationale, and those that required the processing of concepts from different discourse domains. Several of these attributes are particularly interesting in that they suggest that at least some items require a flexibility or fluidity of thought, especially the ability to sort through multiple meanings and to make inferences across domains. Pellegrino and Glaser (1982) argued that ambiguity in the relationship between the A and B terms was one of the major sources of difficulty on the analogy items they examined.

Insert Table 2 about here

Pellegrino and Glaser (1982) propose that subjects solve verbal analogies by first abstracting (i.e., inferring) semantic relationships from the stem of the item, and then evaluating the alternatives in a generate and test mode. But for more difficult problems, they argue that subjects must use information in options to guide the search for the appropriate A-B relationship. Key sources of difficulty, then, are the abstractness or complexity or precision of the rule that must

be generated, and the variability or initial ambiguity in the rules that may be inferred. They hypothesize that this representational variability is a key source of problem difficulty, especially on verbal analogies.

The claim that subjects use information in the alternatives to constrain the inference process on difficult analogies was strongly supported in a study by Alderton, Goldman, and Pellegrino (1985). Alderton et al. (1985) obtained separate estimates for the accuracy of stem processing and alternative processing on both verbal analogy problems (considered here) and verbal classification problems (considered below). The goal was to obtain separate estimates of the probability that subjects used information in the stem to "reason forward" to the answer (which was selected from the set provided), or to "reason backward" from the alternatives to stem. Reasoning from alternatives to stem could lead to a successful solution, in that the alternatives could be used to constrain the search for relationships among terms in the stem, thereby allowing recognition of

an alternative that would not have been selected on the basis of initial stem processing. (Whitley (1980) refers to this as "event recovery.") But analysis of alternatives could also be a source of error, such as when an incorrect but plausible distractor was chosen even though stem processing had been accurate. Previous work with children showed that distraction was a major source of error for children (Goldman, Pellegrino, Parseghian, & Sallis, 1982).

Results showed that recognition was more important than stem processing in predicting overall problem solving success--substantially so in one sample and moderately so in another. Distraction, although generally not as important for adults (4% of items) as for children (36% and 27% of items for third & fifth graders, respectively, in Goldman et al., 1982), was a significant predictor of overall accuracy for low scoring subjects, but not for high scoring subjects. Thus,

the assumption that adults engage in exhaustive encoding of the A-B terms in an analogy may apply only simple problems. For more difficult problems, successful analogy solvers used both feed-forward and feed-backward strategies.

Several other studies confirm the hypothesis that more complex models are needed to characterize the behavior of subjects on difficult items. Gitomer, Curtis, Glaser, and Lensky (1987) studied subject's eye fixation patterns on verbal analogy items that varied widely in

difficulty. Results for easy problems replicated Sternberg's (1977) finding that high-ability subjects spend proportionately more time processing the stem (encoding, inference, mapping) than low-ability subjects. The pattern was reversed on difficult problems. Highs actually spent more time processing stem words after, rather than before, looking at answer options. In addition, highs were much more likely than lows to consider all answer choices on difficult problems. Gitomer et al. (1987) interpret their results as supporting the claim of Bethell-Fox et al. (1984) that explanations of individual differences on analogies of nontrivial difficulty must include an explanation of this differential flexibility of subjects in using different solution strategies as items increase in difficulty. They suggest that the important metacognitive skill may be the ability to monitor the success of each processing step--particularly the crucial step of insuring that the A-B relation maps on to the C-D relation. They caution, however, that processing flexibility may be only one aspect of observed ability differences, given the importance of vocabulary knowledge in predicting difficulty of verbal analogies.

But vocabulary knowledge is a slippery variable. Increasing difficulty by using infrequent words may actually introduce construct-irrelevant variance, at least if the goal is to measure Gf rather than Gc. Horn (1972) showed this in a study in which he presented two types of analogies, which he called esoteric word analogies and common word analogies. Esoteric analogies used infrequent words about a topic unfamiliar to most adults. A test of the former loaded on a Gc factor (along with vocabulary), whereas a test of the latter split its variance between a Gc and a Gf factor. Marshalek (1981) made a similar observation about tests of vocabulary knowledge: correlations with reasoning (Gf) increase as vocabulary items emphasized precise understanding of common concepts, and declined as items required vague understanding of common concepts or knowledge of infrequent words.

Sentence Completion

The sentence completion test is one of the oldest mental tests. In the late 1890s, Ebbinghaus was using sentences from which one or more words had been deleted to study learning

and memory. From this work, Ebbinghaus concluded that the ability to combine information was a major component of intelligence.

As with other verbal tasks, incomplete sentences can be constructed to emphasize different aspects of verbal knowledge and reasoning. Consider the following:

1. Children who do not mind their parents are said to be ____ .

   disrespectful unkind disobedient

2. Yesterday, Tom ____ on a trip to Mississippi.

   go went gone

3. The ice will ____ when the sun comes out.

   freeze melt evaporate

4. I am older than Bob, but Bob is ____ than I am.

   younger shorter taller

The first sentence emphasizes vocabulary knowledge; the second, knowledge of grammar (syntax); the third, general word knowledge; and the fourth, reasoning.

Other than a few isolated studies of cloze tests in reading, sentence completion tests have received little attention as measures of reasoning ability. One notable exception is the work of Buck, VanEssen, Tatsuoka, and Kostin (1998) with sentence completion items on the SAT. Buck et al. (1998) treated sentence completion items as mini-reading comprehension tests. They coded four different types of attributes for each item: vocabulary knowledge in both stem and options (e.g., word length, word frequency); syntactic complexity (e.g., number of words in sentence, use of negation); rhetorical and semantic structure (e.g., two missing words in item, opposite meanings, connections or relationships between ideas); and content (e.g., unfamiliar topics, scientific topics). The final rule space analysis retained 20 of these attributes and three interactions among them. All
of these attributes correlated significantly with total score ($\underline{r}$ = .18 to .62) and had significant beta weights in the regression of total score on attributes. Some of the attributes identify aspects of items that can be directly mapped onto cognitive skills; others require unpacking. Further, although total score is a useful first criterion, relationships between attributes and a more finely differentiated set of reference constructs would be even more informative. For example, do some

attributes show

stronger relationships with reasoning? with verbal fluency? As suggested in the sample sentences above, incomplete sentences can be constructed that emphasize different abilities. Clearly, more work needs to be done, although the Buck et al. (1998) report provides an excellent overview of variables that influence processing on the type of sentence completion items used on the SAT.

Classification Problems

Although classification problems are commonly used as measures of inductive reasoning, they have not been the object of much experimental effort. In a typical classification problem, the subject sees three or more words or figures and must decide which of a set of alternatives belongs in the same set. Sternberg and Gardner (1983) studied problems of the form A B, C D : E, in which the subject's task is to decide whether E fits better with A and B or with C and D. The model for these problems contained the same component processes specified for analogies and series problems (see pp. below). However, classification problems on ability tests typically take a different form. The problem stem contains three or more elements which are alike in some way. The subject's task is to decide which of four or five alternatives also belongs in the set. For example, the stem might present the words: POUND GUILDER FRANC. The stem here appears to be "currencies" or perhaps "foreign currencies" or even "Western European currencies." Since the number of elements in even the most restrictive set is quite large, its is unlikely that the subject will correctly generate the desired answer without examining the alternatives. Analogy items differ in that the set of acceptable alternatives is typically much smaller. Alderton et al. (1985) used the procedures previously described (see pp.___) to estimate the probability that adult subjects inferred the correct relation rule for the stem elements, applied this rule to generate the target alternative, could recognize the correct alternative even if the inferred rule would not have included the keyed answer, and would select an incorrect alternative even if correct rule or category had been generated in stem processing. Thus, four process-outcome scores were estimated for each subject: inference, application, recognition, and distraction.

Results showed that low-scoring subjects were much more likely to be fit by the model that included a component for distraction than were high-scoring subjects. Recognition (weighted positively) and distraction (weighted negatively) were better predictors of total score than were inference and application. Other studies (see Goldman & Pellegrino, 1984) suggest that inference is a more important predictor of success on classification problems for children. Thus, adults seem to make better use of the alternatives to reason about the problem than do children. Classification problems thus differ in interesting ways from analogy and series completion problems. Future studies might investigate performance on figural classification problems (see Figure 1). It would also be interesting to separate the rule or category inferencing step (which is the only operation on tests such as Similarities in the Wechsler [1991] scales) from the recognition process, which is emphasized on verbal classification problems.

Seriation Problems

Seriation problems or linear syllogisms are commonly used in tests of reasoning abilities. Figure 1 shows examples from the CogAT (Thorndike & Hagen, 1993) that use numerical content such as the following:

I.	1 dime + 2 nickels	I is worth more than II.
II.	25 pennies	I is worth less than II.
		I is worth the same amount as II.

Series problems can present any number of elements to be arranged. The quantitative example above uses only two terms; verbal series problems typically use three terms. Difficulty of problems can be predicted from the number of terms, the order of presentation of elements, the use of negations, and the presence of marked adjectives (Wright & Dennis, in press). In particular, items are easier if they have fewer terms, if the first premise refers to the first or last term rather than the middle term (Huttenlocher, 1968), and if the premises are worded positively and use unmarked adjectives
(e.g., “better” rather than “worse”). Wright and Dennis (in press) showed that one could construct

a test of linear syllogistic reasoning of known difficulty by systematically varying these facets of item difficulty.

Theorists differ, though, in explanations of how subjects represent and solve such problems. Huttenlocher (1968) argued that subjects create a visual mental model of the elements. Clark (1969) argued instead that subjects represent premises linguistically and compare them. Johnson-Laird (1972) proposed that subjects use both types of processes: a spatial representation early in problem solution and a linguistic representation later on. Sternberg (1980) also argued that both linguistic and spatial representations are used, but on different problems. Johnson-Laird (1985) seems to agree:

he notes that many experiments suggest both that in processing many linear syllogisms, different subjects employ different strategies and that some subjects can be induced to change their strategies.

Sternberg (1986) argues that the primary sources of individual differences on such problems are encoding and combining the premises. Encoding requires that the subject apply a fairly complex set of procedural rules (e.g., for processing negations, marked adjectives). Combining premises requires the construction of some sort of representation or model in working memory that can be coordinated with the premises as they are encoded. Several working-memory functions are thus required, which include but go considerably beyond simultaneous storage and processing. Working memory requirements can be even more substantial, when premises are presented sequentially and singly (see, e.g., the grammatical reasoning test in Kyllonen and Christal, 1990).

Matrix Tests

Matrix tests--particularly the Progressive Matrices test of Raven (1938-1965)--have long been used as a marker for Gf (see, e.g., Figure 1). Several analyses of this test have been reported (Jacobs & Vandeventer, 1972; Hunt, 1974), including an extensive investigation of Carpenter, Just, and Shell (1990). The Carpenter et al. (1990) report is particularly noteworthy, because the authors used a variety of methods--including task analysis, protocol analysis, and computer

simulations, and an equally diverse array of dependent measures--retrospective reports, eye fixations, response errors and response latencies, and success (or failure) of the simulation programs.

Carpenter et al. (1990) began with a task analysis of the progressive matrices test, which suggested that five different types of rules were used to solve items on the test. Ordered from simple to complex, the rules were (a) identity relations, in which an element is the same across all rows or columns; (b) pairwise progression, in which an element changes systematically from cell to cell (e.g. decreases in size across columns); (c) figure addition or subtraction, in which the first two entries combine to make the last entry; (d) distribution of three, in which an object or attribute appears once in each row or column; and (e) distribution of two relations, in which one of the elements in the distribution of three rule has a null or nonmatching value. Some progressions could be described by more than one rule, and some rules could be described differently. In particular, the difficult "distribution of two values" rule could be represented perceptually as "add (or synthesize) elements of figures, but two identical elements cancel." In problems with multiple rules, subjects must discover which elements in three entries in a row are governed by the same rule--a process Carpenter, Just, and Shell call correspondence finding. In order to do this, subjects must focus on one attribute (shape, number, orientation, shading) and determine if the chosen attribute for each pair of entries in a row matches one of the known rules. To further complicate matters, entries in columns 1 and 2 may follow a rule, but entries in columns 2 and 3 may not. Significantly, Simon and Kotovsky (1963) argued that solving series completion problems also requires correspondence finding, pairwise comparison of adjacent elements, and

the induction of rules based on patterns of similarities and differences. Thus, the first source of individual differences on the matrix test was hypothesized to be the ability to infer these sorts of abstract relations.

Some subjects (12 students) were asked to think aloud while they solved problems and their eye fixations were recorded. Other subjects (22 adults) worked silently and then described

the rules that motivated their response. "The most striking feature of the eye fixations and verbal protocols was the...incremental nature of the processing" (p. 411). In other words, subjects appeared to solve problems by decomposing them into smaller subproblems, which were then solved. Thus, it was hypothesized that the second major source of individual differences was the ability to generate subgoals in working memory, to monitor progress toward attaining them, and to set new subgoals as others are attained.

Simulation models. Hypotheses about the processes used to solve items were tested in two simulation programs: FAIRAVEN, which performed at the level of the median college student in the sample, and BETTERAVEN which performed at the level of the best subjects in the sample.

FAIRAVEN consisted of 121 production rules in 3 categories: perceptual analyses (48%), conceptual analyses (40%), and response generation and selection (12%). The point-biserial correlation between the average error rate for 12 subjects and FAIRAVEN's success or failure

was r (32) = .67, indicating a reasonable correspondence. A comparison with error rates for larger, more representative samples used in norming the test were not reported.

FAIRAVEN limitations were (a) it could not "induce" more complex rules; (b) it had no way to backtrack if a hypothesized correspondence was incorrect and thus, it had difficulty where correspondence was based on texture or location; and (c) too many high level goals at once overwhelmed the concurrent processing of goals.

Insert Figure 8 about here

BETTERAVEN was designed to overcome these limitations. Differences are shown schematically in Figure 8. In particular, it exercised more direct strategic control over its own processes through the addition of a goal monitor. BETTERAVEN also could induce (i.e., "knew") more "abstract" rules, such as the "distribution of two" rule. The goal monitor consisted of fifteen new productions that set main goals and subgoals. The main purposes of the goal monitor were (a) to ensure higher-level processes occur serially (do one thing at a time); (b) to

provide an order for inducing rules (conflict resolution); and (c) to maintain an account of the model's progress toward its goals. With these enhancements, BETTERAVEN performed at the same level as the best college students in the sample. Carpenter et al. (1990) conclude that what matrix tests such as the Progressive Matrices measure 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 decomposition, and the differential ability to form higher level abstractions" (p. 429).

A theory-based matrix test. Embretson (in press) used the Carpenter et al (1990) theory to specify templates for 30 matrix items. Twenty-four templates were identical to item structures studied by Carpenter et al. (1990), although two item structures that contained the "distribution of two" rule were reclassified as solvable by an easier figure combination rule. Six new structures, consisting primarily of pairwise or "distribution of three" relationships, were added. A list of 22 objects (such as square, cross) and seven attributes (such as "increases in size") was combined with the item templates to generate items. Finally, seven distractors were created for each item to contain one or more inappropriate attributes or objects. Items were then administered on a computer.

As expected, the best predictor of item difficulty and response time was a variable that summed the relational level of the rules used to construct the item (i.e., identity = 1, pairwise progression = 2, combination = 3, distribution of three = 4, and distribution of two = 5). This variable was hypothesized to represent working memory load (including executive processes for goal setting and strategy monitoring) but also represents difficulty (or abstractness) of inference. It correlated $r = .71$ with difficulty for 150 generated items. A second study showed that a test composed of 34 of these items correlated $r = .78$ with the 48-item progressive matrices test, and both had similar patterns of correlations with subtests of the Armed Services Vocational Aptitude Battery.

There are many ways one can test the adequacy of a theory of a test, one of which is to use the theory to generate new items. If these items behave like items on the source test, then

confidence is raised in the theory, at least as it describes sources of difficulty that have an impact on individual differences in the construct measured by the test (Embretson, in press; Nichols, 1994). However, identifying sources of difficulty is not the same as identifying individual differences in processes

that are influenced by those variables. A second study by Embretson (1995) illustrates this point.

This study once again examined performance on the 150 matrix items used in the previous study. The goal of this study, however, was to estimate the relative contributions of individual differences in general control processing and in working memory capacity to individual differences in performance on these matrix items. Some have emphasized the importance of executive functions (such as assembling performance programs, monitoring their implementation, etc.) in understanding individual differences in intelligence (e.g. Belmont & Butterfield, 1971; Sternberg, 1977, 1985; Snow, 1981), whereas others have emphasized the role of working memory (Pellegrino & Glaser, 1980; Carpenter et al., 1990). Indeed, Kyllonen and Christal (1990) found that reasoning ability correlated approximately $\underline{r} = .8$ with working memory ability in a series of larger studies. Embretson (1995) attempted to distinguish these hypotheses using a multicomponent latent-trait model. The model posited that two latent variables were responsible for individual differences on the task: working memory capacity (the influence of which varied according to the memory load of the item) and control processing (assumed to be required equally by all items). Thus, although the second factor was interpreted as reflecting control processes, it could reflect these processes or any other processes required by all items. Others have argued that control processes are most strongly engaged when tasks are perceived to be of moderate difficulty (Borkowski & Mitchell, 1987) and thus would not be required equally by all items. In any event, results showed that the two latent variables accounted for 92 percent of the variation in total score, with the "control process" latent variable accounting for more variance than the "working memory" latent variable.

Analyses of matrix problems provide the strongest evidence to date for the importance of control processes for high levels of performance on difficult reasoning tasks. But as Embretson's

(1995) study shows, working memory or attentional resources also play an important role. Further, control and assembly processes are probably best understood as aspects of a working-memory system rather than as alternatives to it (Baddeley, 1996).

Theories of Reasoning that Span Several Tasks

Cognitive psychology--particularly the information-processing branch of it--has been characterized as task-bound (Newell, 1980). This criticism is amply demonstrated in many of the studies reviewed to this point. Most reflect intensive analyses of particular tasks. Although such analyses are a useful first step in developing a theory of reasoning or intelligence, ultimately one must identify commonalities across tasks. There are two aspects to be considered, which are nicely captured in Embretson's (1983) distinction between construct representation and nomothetic span. Construct representation refers to the identification of psychological constructs (e.g., component processes, strategies, structures) that are involved in responding to items on tests. Processes of most interest are those that are common across families of tests that collectively define individual difference constructs such as inductive reasoning ability, or those that fall along one of the spokes

of a radex.

Nomothetic span, on the other hand, concerns the correlates of individual differences on a tests. Of the many processes that are involved in performance on a particular task, only some will be shared with other tasks, and of these common processes, an even smaller subset will be responsible for individual differences that are common across tasks. In other words, even processes and structures that are common to all tests in a family of reasoning tasks may contribute little or not at all to individual differences in inductive reasoning.

With these caveats in mind, then, I examine proposals for theories of reasoning that cut across task boundaries. In the final section, I reconsider hypotheses about which processes and structures are most responsible for individual differences in I = Gf = G.

Pellegrino's summary. Pellegrino (1985; see also Pellegrino & Goldman, 1987) argues that inductive reasoning tasks such as analogies, series completions, and classifications all require

four types of processes: encoding or attribute discovery, inference or attribute comparison, relation or rule evaluation, and decision and response.

Encoding processes create mental representations of stimuli on which various inference or attribute-comparison processes operate, the nature of which vary across tasks. In an analogy, the inference process must determine how various terms (particularly the A and B terms) are related to each other. In classification problems, the inference process must identify a rule or category that is shared by all the terms. In series problems, the inference process must identify the pattern in a sequence of letters or numbers. Inference processes are not sufficient for problem solution, however. The problem solver must also determine relationships among two or more first-order relationships in the problem. In an analogy, for example, the relationship between A and B must be identical to the relationship between C and D. In a matrix problem, the relationship among elements in one row must be the same in the other two rows. Pellegrino (1985) argues that one of the

most important aspects of inductive reasoning is the ability to create two or more of these complex relationship structures in memory and to determine their consistency. Errors occur when working-memory resources are exceeded.

Sternberg-Gardner theory. One of the more ambitious attempts to generalize across analogies, series completion, and classification problems was put forth by Sternberg and Gardner (1983). Analogy items were of the form A:B::C: (D_1, D_2), i.e., A is to B as C is to which of two given D alternatives. Series completion items were of the form A B C ::D: (E_1, E_2), where the series ABC must be carried to D and then extended to one of two given E alternatives. Classification items were of the form A B, C D :E, where subjects decide whether E fits better with class A B or class

C D. The claim here is that a common information-processing model can identified that applies to all three item types.

Sternberg and Gardner argue that solving such items requires seven different component processes: encoding, inference, mapping, application, comparison, justification, and response.

Encoding refers to the process of activating information in long-term memory on the basis of information received through the senses. What is activated depends on the contents and organization of information in memory, as well as on the perceived demands of the task and on residual activation from previously encoded items. Inference refers to the process of discovering relationships between two concepts activated by encoding processes. For verbal analogies, this may be modeled as the attempt to discover whether two concepts are related in semantic memory, and if so, what the nature of that relationship might be. A weak knowledge base might support the inference that two concepts are somehow associated, but might not contain information on the nature of that association. Some relationships seem to be pre-stored, whereas others must be determined by comparing attributes of the terms or by comparing their labeled relationships with other terms in memory. The inference step can also require the finding of relationships between relationships. For example, the pairs up-down and black-white are both opposites, and so the relationship between the relationships is one of identity. Mapping and application are similar. Mapping refers to the process of inferring the relationship between the A and C terms. Application refers to the process of generating a term that is related to C in the same way that A was inferred to be related to B. Comparison or evaluation refers to the process of comparing the internally generated answer (D) to the D options provided to determine which is most nearly correct. If none of the available options meet the individual's criterion for acceptability, then the individual may recycle through some or all of the previous model steps, a process sometimes called justification. True-false analogies in which a single answer option is presented do not require this step. Response is not estimated as a separate component but is assumed to be combined with preparatory and other unspecified sources of variance and reflected in the catchall intercept parameter.

Sternberg and Gardner tested their theory in a series of three experiments in which 18 subjects solved a total of 1,440 analogy, series completion, and classification tasks. All tasks used a common set of animal-name stimuli. Although this limited generalizability, it allowed the authors to use previous multidimensional scalings of similarity judgments on these stimuli to make

predictions about the difficulty of inducing rules in each of the three tasks. Following Rumelhart and Abrahamson (1973) and Henly (1969), it was assumed that the difficulty of inferring relations between concepts is a monotonic function of the similarity between them. To operationalize the measurement of this distance, it was assumed that (a) memory structure may be represented as a multidimensional Euclidean space, and (b) judged similarity between concepts is inversely related to distance in this space. To solve various induction problems, it is assumed that the individual must evaluate the distance between ideal answer and each of the alternatives. The location of the ideal point varies across different item formats as shown in Figure 9. For example, in an analogy problem, the ideal point coincides with the fourth vertex of a parallelogram, whereas for a classification problem, it falls at the centroid of the n terms that define the stem.

Insert Figure 9 about here

Validation of component scores. Individual differences in the speed of performing component processes on one task were expected to show correlations with similarly-named component scores on other tasks, and also with external reference-ability tests, especially reasoning tests. Early studies (Sternberg, 1977) found generally small and inconsistent relationships both among component scores, and between component latencies and reference tests. The strong correlations with reasoning tests came from the preparation-response parameter. However,

samples were often too small for correlational analyses (n = 16 or 24 in each of Sternberg's three experiments). Sternberg and Gardner (1983) did find significant correlations between inference and comparison components and reference reasoning tests. Nevertheless, Brody (1992) criticized the Sternberg-Gardner theory because correlations among component scores showed only weak evidence of convergent and discriminant validity. For example, the average correlation (across tasks) for components with the same label was r = .32, which was only slightly higher than the average correlation among components with different names of r = .24. Once again, though, sample size (n = 18 Yale undergraduate students) cautioned interpretation.

Melis (1997) also examined the correlates of component scores. He administered three figural, three verbal tasks, and six reference tests to 72 undergraduate students. Two experimental tasks emphasized encoding (verbal or figural); two emphasized reasoning (with verbal or figural stimuli); and two, evaluation (again with verbal or figural stimuli). Precuing was used to unconfound component scores. Componential models showed excellent internal validity. Furthermore, a multidimensional scaling of component scores across tasks showed a tendency for like-named components to cluster together. However, only 14 of the 114 correlations between reference tests and component scores were statistically significant, and some of these in the "wrong" direction.

Several other studies have estimated correlations between same-named components on different tasks. In the Alderton et al. (1982) study of verbal classification and verbal analogy, accuracy scores were estimated for both stem and option processing. Significant cross-task

correlations were obtained for both inference scores (r = .47) and recognition scores (r = .42). On the other hand, Whitely (1980) failed to obtain significant correlations among common components for analogy and classification tests. In the spatial domain, Mumaw, Pellegrino, Kail, and Carter (1982) reported both convergent and discriminant validity for component scores on two spatial tasks. Thus, the evidence supporting the convergent and discriminant validity of individual differences in component scores is mixed.

Although such correlations do address the issue of the validity of component scores as individual-difference measures, nonsignificant correlations among similar components should not be viewed as indictments of the information-processing models themselves. The problem (which is discussed on p. ____) is that individual differences that are consistent across trials within a task are removed when component scores are estimated for individuals.

Other evidence can be offered to support the claim of process congruence or identity across tasks. For example, Sternberg and Gardner (1983) compared estimated values for encoding, comparison, and response components across tasks and found them to be equivalent. More important would be the demonstration that components with the same name on two tasks are similarly influenced by the same set of independent variables (see Cronbach, 1957, for an example comparing "treatments" and Pieters, 1983, for a discussion of process congruence and independence in the additive factors method).

The important point for this discussion, however, is that degree of correlation between component scores and reference tests is less important than is the identification of information-processing models that describe how subjects solve tasks. Validation of individual-difference measures might better proceed through subtask scores that do not require subtraction.

Developmental differences. Scattered other studies also suggest processes that are common across tasks. For example, developmental changes in analogical reasoning have been studied by several investigators with interesting results. In one study of pictorial analogies, eight-year-olds appeared not to use a mapping process (Sternberg & Rifkin, 1979). In another study using verbal analogies, preadolescents appeared to shift strategies as item complexity increased,

changing from
an analogical reasoning strategy to a strategy in which responses were chosen on the basis of their associative relatedness to the C term. Adolescents and adults continued to use analogical reasoning even on the more difficult items. Heller (1979) reported similar ability-related differences among high-school students, with low-verbal subjects more often using the associative strategy. Goldman, Pellegrino, Parseghian, and Sallis (1982) also found that older children (ten-year-olds) were less likely to be distracted by foils that were associates of the C term than were younger children
(eight-year-olds). Retrospective reports suggested even more substantial differences in processing strategy. Older children were more likely to understand that the C to D relationship had to mirror the relationship between A and B. Pellegrino (1985) argues that younger and less-able students have particular difficulty in remembering and comparing multiple relationships, possibly because they do not understand the "rules of the game" or because problem complexity exceeds mental resources.

Distractors also function differently for adults and children, and for subjects of high and low ability. Bethell-Fox et al. (1984) found that four alternative items like those found in many mental tests were solved differently than otherwise similar two-alternative items. Other data (see Pellegrino, 1985, p. 212) suggest that high ability subjects are better able to recognize the correct answer on a forced-choice analogy test even when they have processed the stem incorrectly. Snow (1980b) and Whitely and Barnes (1979) report similar evidence for subjects working backwards from the options provided. By combining analyses of eye fixation and componential models of latencies and errors, Bethell-Fox et al. (1984) also found evidence that lower ability adolescent subjects shifted strategies on difficult items, changing from one in which they constructed a response and compared it to the alternatives to one in which they attempted to eliminate response alternatives. High-ability subject, however, showed little evidence of strategy shifting, probably because most items were relatively easy for them.

Sternberg's Unified Theory of Reasoning. Sternberg (1986) claims that there are three kinds of reasoning processes, any one of which define a task as a reasoning task. The three processes are (a) selective encoding (distinguishing relevant from irrelevant information), (b) selective comparison (deciding what mentally stored information is relevant for solving a problem), and (c) selective combination (combining selectively encoded or compared information in working memory). Furthermore, the three processes define a reasoning situation only to the extent that they are executed in a controlled rather than in an automatic fashion. This implies that the extent to which a task measures reasoning depends on the relative novelty of the task for the individual.

These processes are implemented by various sorts of inferential rules. Procedural rules include operations called performance components in earlier theories (e.g., inference, mapping, application). Declarative rules vary by problem content and specify the type of semantic relations allowed in a problem. (For verbal analogy problems, for example, the set of possible semantic relations includes equality, set-subset, set-superset, static properties, and functional properties). Not all rules are rules of reasoning; reasoning rules are those that serve the functions of selective encoding, selective comparison, and selective combination. Thus, mnemonic strategies and computation algorithms are not reasoning rules.

The theory also claims that the probability that particular influential rules will be used in the solution of a reasoning problem and will be influenced by mediating variables, such as the individual's subjective estimate of the likelihood of the occurrence of a rule, the individuals' prior knowledge, working memory capacity, and ability to represent certain types of information (e.g., spatial versus linguistic).

Sternberg claims that the major difference between inductive and deductive reasoning is that the difficulty of the former derives mainly from the selective encoding and comparison processes, whereas the difficulty of the latter derives mainly from the selective combination process. Thus, for verbal analogies, the primary difficulty is determining which of the many features of the A term are relevant to the B term as well. For example. in the analogy

paper:tree::plastic:?, one must decide which of the many attributes of the word paper (that we write on it, that it sometimes comes in tablets, that printers use it, that it is a short form of the word "newspaper," that it is made from wood, etc.) also overlap with what one knows about the word tree. In contrast, figural analogies tend to emphasize selective encoding. A key difficulty of problems such as those shown in Figure __ is deciding which features of the stimuli to attend to in the first place.

Series completion problems not only require many of the same processes as analogies (Greeno, 1978; Pellegrino & Glaser, 1980; Sternberg & Gardner, 1983), but also emphasize selective comparison. In a typical series problem, there are many possible relations that could be obtained between successive pairs of numbers or letters. For example, in the series 1, 3, 6, 10,..., the relation between the first two digits could be plus 2, times 3, next odd number, etc. The relation between 3 and 6 could be plus 3, times 2, etc. Problem difficulty is highly related to the obscurity

of the rule. However, when multiple rules account for a series, the "best" rule is typically the most specific rule. A similar set of arguments apply to the analysis of classification problems.

For deductive reasoning tasks such as categorical syllogisms, however, the main source of difficulty is not in encoding the terms or even in selectively comparing relations among them, but rather in keeping track of the ways in which terms can be combined. Consider, for example, a categorical syllogism such as "Some A are B. All B are C." Is the conclusion "Some A are C" valid? Information processing models of syllogistic reasoning all share four stages of information processing, which Sternberg (1986) calls encoding, combination, comparison, and response. In the encoding stage, the individual must create a mental representation of each premise that is amenable to mental transformation. Some have argued that this representation is imagistic or at least like a Euler diagram (Erickson, 1978). Others claim it is propositional (Johnson-Laird & Steedman, 1978). Regardless of the nature of the representation, the important aspect of this encoding is that it is not particularly selective. In other words, all "All A are B." statements should be translated into the same representation, regardless of their context. Indeed,

there are four forms of statement permissible for each premise and the conclusion, and four different sets of mental representations, one for each of these four forms. However, for inductive reasoning problems such as series completions, classifications, or analogies, there are a vast number of possible representations.

It is the large number of combinations between representations of premises that taxes processing resources. For example, the problem "Some B are C. Some A are B." involves 16 combinations (four for each of the two premises). Further, the exact inferential rule used also appears to be a major source of difficulty, although there is controversy as to exactly what these rules are. For example, Erickson (1974) specifies rules for mapping premises on to Euler diagrams, but claims that the mapping is stochastic rather than deterministic. More important, however, has been the recurring finding that many other factors (what Sternberg calls mediators) influence performance as categorical syllogisms. For example, subjects show flagrant biases in solving such problems as a function of the emotionality of the premise, subjects' agreement with the content of the premises, abstractness of the content, and even the form in which the problems are presented. This suggests that, although such problems may be interesting candidates for research, they are probably not good candidates for assessments of individual differences in reasoning abilities.

Another type of deductive reasoning task that has been extensively studied is the linear syllogism. These are problems of the sort "Bill is taller than Mary. Mary is taller than Sue. Who's tallest?" On tests of inductive reasoning, problems of this sort have anywhere from two to four terms, with the most typical number being three.

Different theories of how such problems are solved have been proposed. Clark (1969) argued that solvers code the premises linguistically, Huttenlocher (1968) argued that premises were coded spatially, and Sternberg (1980) argued that both types of representation were used. As in other deductive reasoning problems, the major source of difficulty is not in encoding the terms or in comparing them (for example, to know that "short" is the opposite of "tall"), but rather to combine the information in the premises into a single mental model. Unlike linear syllogisms,

however, there are fewer content-induced biases to cloud performance. Indeed, the most likely bias occurs when the premise contradicts one's personal knowledge, such as when one knows that Mary is shorter than Sue, whereas the problem asks one to envision the opposite. Such contrafactual reasoning can be deliberately introduced into problems (e.g., "imagine that mice are larger than elephants," etc.).

Verbal and Spatial Abilities

Although reasoning or Gf abilities are the focus of this chapter, any discussion of complex information processing and intelligence would be incomplete without at least some overview of the research on verbal and spatial abilities. One might also discuss quantitative or symbolic reasoning abilities as well, although individual differences in such tasks are often completely subsumed in the Gf factor. This is not the case for verbal and spatial abilities. Both a broad verbal factor (Gc) and

a broad spatial factor (Gv) can be identified that are at least somewhat independent of Gf. The objectives of this section, then, are to explore both the overlap and the uniqueness, i.e., to explain (a) why verbal comprehension or spatial visualization are primarily measures of G or Gf and (b) why each measures something unique.

Verbal Abilities

Verbal abilities are central to all theories of intelligence. Spearman (1927) notes that a verbal factor was one of the first to be identified among the residual correlations in a matrix after G had been removed. Thurstone (1938) identified two factors that he called Verbal Relations, which roughly corresponded to the factor Spearman had identified, and a second factor that he called

Word Fluency. This distinction, or something like it, persists in modern summaries. Carroll (1983) claims that the major factor distinction within the verbal domain is between Gc (general crystallized verbal abilities) and Gi (general idea production). Thus, the first task for a cognitive theory of verbal abilities is to explain this distinction. The second task is to explain the high correlation between general reasoning abilities (G or Gf) and verbal comprehension and

vocabulary test. The overlap is more understandable for comprehension tests than for vocabulary tests, which seem to represent little more than the number of words the examinee has learned through instruction or exposure.

Verbal comprehension versus word fluency. Verbal comprehension abilities are typically assessed by tests of oral and written comprehension, sentence completion, verbal information, and vocabulary (including defining words, selecting or generating synonyms and antonyms). Verbal fluency, on the other hand, refers to the speed and ease with which ideas, words, sentences, and other linguistic responses can be generated, often on the basis of perceptual cues such as the first letter of a word or its final phoneme. It is not one but a loose collection of many abilities (Carroll 1941, 1993).

The same task can require both fluency and comprehension abilities. For example, Janssen, DeBoek, and Vander Steene (1996) showed that an open synonym task, in which participants were required to generate synonyms for a given word, could be decomposed into a generation component and an evaluation component. Individual differences in a generation subtask were primarily related to verbal fluency, whereas individual differences on an evaluation task were primarily related to verbal comprehension.

Word fluency abilities would seem to be dependent upon verbal comprehension abilities: words must be understood before they can be generated (Sincoff & Sternberg, 1987). On the other hand, the literature on savants provides examples of individuals who can retain and reproduce much oral language while understanding little or none of it.

Verbal comprehension tests emphasize semantic processing and show high correlations with general reasoning measures. Verbal tests that show relatively low correlations with reasoning are more diverse. Such tasks are usually simpler; emphasize speed of response; require knowledge of linguistic conventions (e.g., grammar); require phonological or articulatory processing; or require memory for the exact order of events, words, letters, or sounds. Conversely, correlations between verbal tasks and reasoning increase as tasks require (a) more inferencing; (b) the integration of information (across words, sentences, or paragraphs; or with prior knowledge); (c) knowledge of abstract words; or (d) precise understandings of common words and concepts. Sex differences are also inversely related to the correlation of the task with G. The female advantage is large on many of the more specific verbal tests and factors, but small or nonexistent on verbal comprehension and vocabulary tests.

Research in the information-processing paradigm has suggested several hypotheses about this contrast. One possible difference between tasks that show high correlations with reasoning and tasks that define more specific verbal abilities such as fluency or spelling may be the nature of the memory code on which performance is based. Verbal comprehension and vocabulary tests may estimate facility in creating, retaining, and transforming information coded in a way that preserves meaning (e.g., abstract propositions). More specific verbal (and spatial) measures

may estimate both how well and how fast one can create, retain, and transform information coded in one of several perception-based codes. A string code would preserve order information for sounds, letters, words, and other stimuli. An image code would preserve configural information for spatial arrays. Specific phonological, articulatory, or other codes seem highly likely as well, but have not been much studied (see Anderson, 1983, 1985).

A second hypothesis is that more highly G-loaded tasks place greater demands on working memory (Kyllonen & Christal, 1990). On this view, verbal tasks such as comprehension and inferencing are more highly G-loaded than tasks such as rhyming and spelling because they place greater demands on working memory, particularly for simultaneous storage and processing of information, but also on monitoring and inhibition functions. This leads directly to the second problem: the overlap between measures or reasoning, comprehension, and vocabulary.

Reasoning, comprehension, and vocabulary. The relationship between reasoning ability and verbal ability is perhaps best understood by examining how people understand text that contains unfamiliar or ambiguous words, or conversely, how they come to understand new words.

Word meanings are learned in two ways: by explicit definition and through contextual inference, i.e., through the process of inferring the meaning of a word from context (Werner & Kaplan, 1952; Johnson-Laird, 1983). Daalen-Kapteijns and Elshout-Mohr (1981) used verbal protocols to study the process of abstracting word meanings from contexts. They proposed that subjects first generate a schema or hypothesis for the meaning of an unfamiliar word based on their best interpretation of its meaning in the first sentence in which it occurs. The schema has slots that can then either be confirmed or contradicted by new evidence. This can lead to adjustment or complete reformation of the schema. Ideally, the initial schema would be sufficiently well-articulated to permit an active search for information to confirm it. In their research, low-verbal subjects were less likely to use this strategy of schema-guided search, possibly because they did not have or were not able to activate the appropriate knowledge schema for the neologisms.

Rumelhart (1980) proposed an almost identical model for the process of comprehending written prose. In a series of experiments, Rumelhart presented subjects with a series of somewhat ambiguous sentences describing an event. After each sentence, subjects were asked to tell what they thought the passage was about. Subjects typically had multiple hypotheses about the meaning of the passage after reading the first sentence. Most entertained different hypotheses as new sentences were presented, and eventually all inferred the intended scheme.

Although Rumelhart (1980) did not investigate individual differences in his experiments, Frederiksen (1981) has found that subjects differ in the extent to which they use contextual cues when reading; skilled readers prime a wider range of relevant concepts in memory for a given set of contextual cues than do less-skilled readers.

Sternberg and Powell (1983) have also presented a process theory of verbal comprehension that is based on learning from context. Their theory has three parts: context cues, mediating variables, and processes of verbal learning. Context cues are hints about the meaning of the unknown word contained in the passage. Mediating variables (e.g., "variability of context") attempt to specify how contextual cues may help or hinder the inference process in any particular situation. The theory also hypothesizes three verbal learning processes: selective encoding, selective combination, and selective comparison (see Sternberg & Powell, 1983, for details; also Sternberg, Powell, & Kaye, 1983). Selectivity in encoding, combination, and comparison in part reflects the contribution of well-structured knowledge. Experts in medical diagnosis, for example, selectively attend to certain symptoms because their thinking is guided by their declarative knowledge of diseases (see Lesgold, 1984). Nevertheless, declarative knowledge cannot be the only difference between high- and low-ability subjects, since one then could not explain how some individuals acquire more precise and useful knowledge in the first place. One possibility is that learners differ in their use of certain metacognitive or performance processes when learning, such as systematically testing alternative interpretations in unfamiliar situations, that then lead to a richer and more usefully organized knowledge base to guide new learning (see, e.g., Robinson & Hayes, 1978).

Parts of the Sternberg-Powell theory were tested in an experiment in which subjects were asked to read passages containing one to four extremely low-frequency words. These words were presented with different frequencies and with different contextual cues. Students then attempted to define the words, and their definitions were rated for quality. These ratings were then regressed on independent variables coded to estimate variables hypothesized in the model. When averaged over four types of passages, the model variables accounted for 67% of the variance in the ratings. Rated quality of definition for individuals correlated approximately $\underline{r}$ = .6 with IQ, as well as vocabulary and reading comprehension tests. This is consistent with Marshalek's (1981) claim that the ability to infer word meanings from the contexts in which they occur is the cause of high correlation typically observed between vocabulary and reasoning tests.

Sternberg and McNamara (1985) attempted to incorporate these and other findings into a model of verbal comprehension. Their model includes both the representation of word meaning and the nature and speed of word processing in verbal comparison tasks. They showed that overall latency scores on their tasks correlated significantly with a combination of rate and comprehension scores from conventional reading comprehension tests. The multiple correlation was not strikingly high ($\underline{R}$ = .47), but samples were small.

Their two-stage componential model first compared alternative possibilities for word representation by contrasting defining attribute models and characteristic attribute models. In the former, word meaning depends on a stored list of semantic features that are both necessary and sufficient for a word to refer to an object or concept. In the latter, attributes are neither necessary nor sufficient but rather combine to characterize the referent as in family-resemblance models of concepts (Rosch & Mervis, 1975). The best fit to data was obtained with a mixed model that assumed both defining attributes and a weighted sum of characteristic attributes specifying a word referent. Then, various alternative models were compared to determine whether processing of word attributes and/or answer options was performed in an exhaustive or self-terminating mode.

Modeling suggested that subjects always process answer options exhaustively, that is, they search all answer options given, and apparently also always use both defining and characteristic information about the stimulus words in doing so.

The Sternberg-McNamara theory thus supports a view of verbal ability that posits the acquisition of word meanings from contexts using contextual and other cues to guide selective encoding, combination, and comparison processes to produce schemata. These, in turn, drive further search and hypothesis testing processes. Such schemata have slots for both defining and characteristic attributes of words, and are used exhaustively in tasks requiring the demonstration of verbal comprehension.

To recapitulate, the high correlation between vocabulary knowledge and reasoning seems to reflect primarily the fact that word meanings are generally inferred from the contexts in which they are embedded. But there is a synergism here in that vocabulary knowledge allows comprehension and expression of a broader array of ideas, which in turn facilitate the task of learning new words and concepts. Thus, language functions as a vehicle for the expression, refinement, and acquisition of thought. The humble vocabulary test masks an enormous amount of reasoning and remembering.

Spatial Abilities

Spatial ability may be defined as the ability to generate, retain, retrieve, and transform well-structured visual images (Lohman, 1988). It is not a unitary construct. There are, in fact, several spatial abilities, each emphasizing different aspects of the process of image generation, storage, retrieval, and transformation.

Reviews of factor-analytic studies of spatial abilities (Anderson, Fruchter, Manuel, & Worchel, 1954; Carroll, 1993; Lohman, 1979; McGee, 1979), show a broad array of spatial factors that can be organized hierarchically. A general visualization factor (Gv or Vz) appears at or near the top of this hierarchy (Carroll, 1993; Horn, 1976), usually defined by complex spatial tests such as paper form board, paper folding, and surface development (see Ekstrom et al., 1976). However, the Vz factor is often difficult to separate from a Gf, or reasoning, factor (Guilford & Lacey, 1947; Lohman, 1979; Marshalek, Lohman, & Snow, 1983). Indeed, Vz tests appear to be primarily

measures of G or Gf, secondarily measures of task-specific functions, and thirdly measures of something that covaries uniquely with other Vz tests. A spatial orientation (SO) factor can sometimes be distinguished from Vz. SO tests require subjects to imagine how an array would appear from a different perspective and then to make a judgment from that imaged perspective. A speeded rotation (SR) factor emerges if two or more simple, highly speeded rotation tasks are included in a test battery. Complex, three-dimensional rotation tasks (such as the Vandenberg-Kruse [1978] adaptation of the Shepard-Metzler [1971] figures) generally split their variance between SR and Vz factors. Other distinguishable visual-spatial factors include flexibility of closure, speed of closure, perceptual speed, visual memory, serial integration, and kinesthetic (see Carroll, 1993; Lohman, 1988).

As with verbal abilities, the major tasks for understanding spatial abilities are to explain (a) the systematic individual difference variance that is uniquely spatial an (b) the much larger portion of the vartiation on such tasks that is shared with general reasoning abilities.

Individual Differences in Spatial Cognition. Cognitive psychology has contributed importantly to understanding the uniquely spatial variance on figural tasks through its investigations of how subjects encode, remember, and transform visual images. Seminal research here was that of Roger Shepard and his students on mental rotation (see Shepard & Cooper, 1982, for a summary). The basic finding was that the time required to determine whether two figures could be rotated into congruence was a linear function of the amount of rotation required. On the basis of this and other evidence, Shepard claimed that mental rotation was an analog process that showed a one-to-one correspondence with physical rotation. The second claim was that this rotation process was performed on a mental representation that somehow preserved information about structure at all points during the rotation transformation.

However, most theorists argue that spatial knowledge can be represented in more that one way. One representation (sometimes called an image code) is thought to be more literal (Kosslyn, 1980) or at least more structure- or configuration-preserving (Anderson, 1983). This is the sort of representation that Shepard thought necessary to explain mental rotation. Another

representation is more abstract and is more meaning- or interpretation-preserving (Kosslyn, 1980; Anderson, 1983; Palmer, 1977) and is usually modeled by the same propositional structures used to represent meaningful verbal knowledge. Some of the confusion in understanding spatial abilities can be traced to whether spatial abilities are restricted to image-coded memories and the analog processes that operate on them or whether proposition-coded memories and the general procedural knowledge that operate on them are also considered part of the term.

Although research and theory in cognitive psychology and artificial intelligence suggest much about the nature of spatial knowledge and processes, it does not explicitly address the source of individual differences in spatial processing. Research on this question has followed four hypotheses: that spatial abilities may be explained by individual differences in (a) speed of performing analog transformations, (b) skill in generating and retaining mental representations that preserve configural information, (c) the amount of visual-spatial information that can be maintained in an active state, or (d) the sophistication and flexibility of strategies available for solving such tasks.

The most popular hypothesis has been that spatial abilities may be explained by individual differences in the speed with which subjects can accurately perform analog mental transformations, particularly rotation. However, correlations between estimated rate of rotation and spatial ability vary from highly negative (e.g., Lansman, 1981) to moderately positive (e.g., Poltock & Brown, 1984). Correlations are generally somewhat higher for three-dimensional rotation problems than

for two-dimensional problems (Pellegrino & Kail, 1982; Cooper & Regan, 1982), and for practiced than for nonpracticed subjects (Lohman & Nichols, 1990). However, even moderate correlations between the slope measure and other variables are difficult to interpret. The slope is heavily influenced by the amount of time taken on trials requiring the most rotation. Some subjects make more than one attempt to rotate stimuli on such problems. Therefore, at least for these subjects, the slope better reflects the number of attempts made to solve a problem or simply the time taken

to solve these most difficult problems, and not rate of rotation per se. More importantly, the slope measure ignores individual differences on the task that are consistent across different trial types. These individual differences are captured in the mean or in the intercept scores (Lohman, 1994). On the other hand, correlations between overall-error rates and spatial reference tests are often quite high. Indeed, although the rate of information processing on rotation tasks and accuracy levels achieved under liberal time allotments are necessarily confounded, differences between high- and low-spatial subjects are much greater on the accuracy score than on a rate of information-processing score (Lohman, 1986). One interpretation of this finding is that the amount of information that can be maintained in an active state while it is being transformed is more important than the rate of executing that transformation in accounting for individual differences in spatial ability.

The second hypothesis is that high spatial subjects have sperior spatial working memory resources. Baddeley's (1986) model of working memory hypothesizes a central executive and two slave systems: an articulatory loop and a visual-spatial scratch pad. Perhaps high-spatial subjects can maintain more information in this scratch pad. Kyllonen's (1984) study of ability differences in types and number of errors made on a paper-folding task supports this hypothesis. His study showed that high- and low-spatial subjects differed not so much in the type of error committed but in the number of errors committed. Because of this, Kyllonen (1984) concluded that the main difference between the performance of high- and low-spatials on a paper-folding task was that lows were more likely to forget a fold and then either not perform it or substitute an incorrect fold for the forgotten one. Other theorists (e.g., Just & Carpenter, 1992) emphasize the tradeoff between storage and transformation functions in a unitary working-memory system. By this account, mental-rotation problems are good measures of spatial ability because they place substantial demands on both storage and transformation functions, and require subjects to manage the tradeoff between them. Other evidence in support of this view comes from verbal problems that require imagery for their solution, such as Binet's "It is 12:15. If we switch the

hands on the clock, what time will it be?" Tests constructed on such problems often show high predictive validities, but are not factorially pure (Ackerman & Kanfer, 1993).

The third hypothesis focuses on the nature of the representation. Several investigators have sought to determine whether high- and low-spatial subjects differ in the type of mental representations they create (see, e.g., Cooper, 1982; Lohman, 1988). Individual differences in memory for random forms shows no relationship with performance on other spatial test (Christal, 1958). Thus, it is not so much the ability to remember stimuli but the ability to remember systematically structured stimuli that distinguishes between subjects high and low in spatial ability. Low-spatial subjects seem to have particular difficulty in constructing systematically structured images. High-spatial subjects appear to be able to construct images that can be compared holistically with test stimuli. Differences between high- and medium-spatial subjects are often small in this respect. It is the very-low-spatial subjects who appear qualitatively different (Pellegrino & Kail, 1982; Lohman, 1988).

The fourth explanation emphasizes the role of strategies. It has long been noted that spatial tasks may be solved in more than one way, with some strategies placing greater demands on analog processing that others. Several studies have now shown that the strategies subjects employ on form-board tasks are systematically related to their ability profiles (Kyllonen, Lohman, & Woltz, 1984; Lohman, 1988). The major distinction is between spatial and nonspatial strategies. Subjects using the spatial strategy remember complex polygons by decomposing them into simpler geometric shapes. When they are required to assemble figures mentally, their performance is more influenced by the characteristics of the to-be-assembled figure than by that of the component figures. Time to perform this assembly operation is usually negatively correlated with reference spatial tests. On the other hand, subjects using the nonspatial strategy try to remember complex polygons by associating the figure with another concrete, easily labeled object. When they are asked to assemble figures mentally, their performance is strongly influenced by the complexity of the component figures rather than that of the to-be-assembled figure. Further time to perform the assembly often shows higher correlations with tests of verbal ability than with tests of spatial ability.

Rotation tasks are also solved in different ways by different subjects. Bethell-Fox and Shepard (1988) found that rotation times for unfamiliar stimuli were generally influenced by the complexity of the stimulus. With practice, most subjects learned to rotate all stimuli at the same rate. However, some subjects continued to show effects for stimulus complexity even after much practice. Bethell-Fox and Shepard (1988) argued that these subjects rotated stimuli piece by piece, whereas after practice others rotated them holistically. Carpenter and Just (1978) argued that even practiced subjects do not rotate an image of an entire three-dimensional object, but rather only a skeletal representation of it. In experiments on a cube-rotation task, they found that subjects used different strategies, presumably related to the coordinate system the subject adopted. Low-spatial subjects appeared to rotate the cube iteratively along standard axes, whereas high-spatial subjects were able to use the shorter trajectory defined by a single transformation through both axes (Just & Carpenter, 1985).

Thus subjects of different ability levels and profiles often solve spatial tests in predictably different ways. However, flexibility of strategy in solving such tasks seems to be more related to G or Gf than to spatial ability (Kyllonen et al., 1984). Indeed, subjects high in spatial but low in verbal abilities have been found to apply the same "spatial" strategy to all problems. Perhaps they have no need to switch to other strategies.

Spatial Ability and Gf. In addition to explaining the uniquely spatial variation on spatial tasks, a theory of spatial ability must also explain the substantial overlap between spatial and reasoning abilities. These explanations have taken several forms. Some have argued that the overlap stems from the fact that most complex spatial tests can be solved at least in part through the application of reasoning abilities. For example, when a piece of paper is folded in half, then in half again, and a hole is punched through the quarter-folded paper, one need not envision the unfolding process to infer that the unfolded paper must have four symmetrically arrayed holes, one of which must coincide with the hole that was punched. Such reasoning is facilitated when tasks show where and how the folds are made. A second explanation focuses on the nature of the mental representation. Much spatial knowledge seems to be represented in the same

propositional format that some argue is also used to represent abstract verbal knowledge. Both Palmer (1977) and Kosslyn (1980) claim that only a subset of spatial knowledge is represented in a more literal code that is processed analogically. It is the latter that represents the uniquely spatial knowledge and processes, whereas the former are common to all meaning-making tasks. A third explanation for the overlap focuses on working-memory demands. Spatial tasks are process intensive, thereby placing substantial demands on working memory for the simultaneous storage and processing of information. If reasoning ability is "little more than working memory" (Kyllonen & Christal, 1990), then spatial tasks would be little more than reasoning tasks. A fourth explanation also involves working memory but emphasizes the coordination of different mental models. On this view, individuals often construct different types of mental models when listening, reading, or thinking. For example, Kintsch and Greeno (1985) argued that individuals typically must create and coordinate two types of mental models when solving word problems in mathematics. One model is more text or proposition based, and the other is more image or situation based. Understanding requires not only the construction, but also the active coordination of the two models. The finding that verbally presented spatial problems (such as the paper-folding problem previously described) load on both Gv and Gf supports such an interpretation.

Complexity Continuum Revisited

In their discussion of evidence from correlational studies concerning the nature of G, Snow et al. (1984) argue that the complexity continua in the radex is an important--if not the most important--feature to be explained by a theory of intelligence. Tests that load heavily on the G or Gf typically fall near the center of the radex, whereas seemingly simpler tasks are distributed around the periphery. Moving from the periphery to the center, one steps from task to task along spokes where tasks sometimes seem to build systematically on one another and sometimes increase in complexity in less obvious ways. But what is the nature of this complexity? We are now in a position to examine this question.

Several hypotheses have been advanced to explain how processing complexity increases along the various spokes that run from periphery to G: (1) an increase in the number of component processes; (2) an accumulation of differences in speed of component processing; (3) increasing involvement of one or more critically important performance components, such as inferencing processes; (4) an increase in demands on limited working memory or attention; and (5) an increase in demands on adaptive functions, including assembly, control, and monitor functions. Clearly these explanations are not independent. For example, it is impossible to get an accumulation of speed differences over components (Hypothesis 1) without also increasing the number of component processes required (Hypothesis 2). In spite of this overlap, these hypotheses provide a useful way to organize the discussion.

More Component Processes

Even the most superficial examination of tasks that fall along one of the spokes of the radex reveals that more central or G-loaded tasks require subjects to do more than the more peripheral tests. Many years ago, Zimmerman (1954) demonstrated that a form-board test could be made to load more on perceptual speed, spatial relations, visualization, and reasoning factors, in that order, by increasing the complexity of the items. Snow et al.'s (1984) reanalyses of old learning-task and ability-test correlation matrices showed similar continua. Spilsbury (1992) argues that the crucial manipulation here is an increase in the factorial complexity of a task. However, increases in the number or difficulty of task steps beyond a certain point can decrease the correlation with G (Crawford, 1988; Raaheim, 1988; Swiney, 1985). For example, Crawford (1988) reported a similar in correlation with G for a mental counting task as task complexity increased. Correlations were slightly higher for 8-9 element items than for 11-12 element items. Although both string lengths are supra-span for most individuals, the latter are probably supra-span for all unpracticed participants. Thus, one does not automatically increase the relationship with G simply by making problems harder, or even by increasing the factorial complexity of a task--unless, of course, the added dimension reflects a type of processing or capacity limitation that is central to G. Indeed, there are many hard problems (e.g., memorizing lists of randomly

chosen numbers or words) that are not particularly good measures of G. Furthermore, even for problems that do require the type of processing that causes the test to measure G, problems must be of the appropriate level of difficulty for subjects, or in what Elshout (1985) has called the "zone of tolerable problematicity" (see pp. ____).

Speed or Efficiency of Elementary Processing

This hypothesis has taken several forms. Since this work is reviewed in detail elsewhere in this handbook (____), only a few points will be made here. In its strongest form, the assertion has been that individuals differ in the general speed or efficiency with which they process information (Jensen, 1980, 1982, 1987, 1998). In principle, processing speed could be estimated on any elementary cognitive task that minimizes the import of learning, motivation, strategy, and other confounding variables. Although disattenuated correlations between RT and G can be substantial when samples vary widely in ability (even, for example, including mentally retarded participants), samples more typical of those used in other research on abilities yield correlations between RT and G in the $r = -.1$ to $r = -.4$ range (Jensen, 1982; Roberts, 1995; Sternberg, 1985; Deary & Stoush, 1996). Furthermore, response latencies on many tasks show a pattern of increasing correlation with an external estimate of G as task complexity is decreased. In other words, response latencies for simpler tasks typically show higher correlations with G than do response latencies for more complex tasks. But this is unsurprising. The more complex the task, the more room there is for subjects to use different processes or even to be inconsistent in the execution of a common set of processes. Indeed, for the most complex task, latency becomes more a measure of persistence than ability, which is instead reflected in the quality of the response given.

Others argue that more-able individuals refresh memory traces more rapidly and thus are better able to hold information in working memory--especially when required to effect transformations on that information as well. This is a good example of the inevitable tradeoff between process and structure in cognitive models, because speed of refreshing memory traces would generally be

indistinguishable from the strength or persistence of those traces. Put differently, what one theorist might interpret as a speed of processing difference might be interpreted as a processing capacity difference by another theorist.

In its weak form, the hypothesis has been that although speed of processing on any one task may be only weakly correlated with more complex performances, such small differences cumulate over time and tasks. Thus, Hunt, Frost and Lunneborg (1973) noted that although latency differences in the retrieval overlearned name codes correlated only $r = .3$ with verbal ability, such small differences on individual words cumulate to substantial differences in the course of a more extended activity. Detterman (1986) emphasized the cumulation across different component processes rather than across time. He showed that although individual component processes were only weakly correlated with G, their combined effect was more substantial.

Although individual differences in speed of processing are an important aspect of G, G is more than rapid or efficient information processing. Furthermore, the strength of the relationship between speed of processing and G varies considerably across domains, being strongest ($r \approx -.4$) in verbal domain and weakest ($r \approx -.2$) in the spatial domain. Indeed, for complex spatial tasks, the speed with which individuals perform different spatial operations is usually much less predictive of overall performance than the richness or quality of the mental representations they create (Lohman, 1988; Salthouse, Babcock, Mitchell, Palmon, & Skovronek, 1990).

More Involvement of Central Components

If G is not simply a reflection of more or faster processing, might it be the case that G really reflects the action of particular mental processes? Spearman (1927) was one of the first to argue for this alternative. For him, the essential processes were the "eduction of relations," which Sternberg calls inference, and the "eduction of correlates," which Sternberg calls mapping and application. Evidence favoring this hypothesis is substantial. A common characteristic of tests that are good measures of Gf--such as the matrices, series completion, analogies,and classification

reviewed in this chapter--is that they are all measures of reasoning, particularly inductive reasoning. Many school learning tasks, particularly in science and mathematics, bear formal similarity to Gf tests. Greeno (1978) refers to such tasks, collectively, as problems of inducing structure. Indeed,

the problem of inducing structure in instruction is probably why reasoning tests correlate with achievement tests (Snow, 1980a). But to describe the overlap in this way is not to explain it.

Evidence supporting the hypothesis that particular component process are central to G has been surprisingly difficult to obtain. Sternberg's (1977) investigations of analogical reasoning found little generalizability of component scores for inference across tasks, and at best inconsistent correlations with reference reasoning tests. Rather it was the intercepts that showed more consistent correlations with reference abilities. We now know that this was in large measure an inevitable consequence of the way component scores are estimated (Lohman, 1994). Individual differences that are consistent across items that require different amounts of a particular component processes will appear in the intercept rather than in the component scores. Put differently, if examinees who are fast in making easy inferences are also fast in making more difficult inferences, then most of

the reliable individual difference variance in speed of making inferences will be removed through subtraction. In addition, if those who are faster making inferences are also faster in general in solving items, then these differences will be reflected in the intercepts. Indeed, the conditions under which component scores will show strong and consistent correlations with other variables are exactly the same as those which lead to more reliable difference scores: low correlation between the two scores that are subtracted and an increase in variance across scores. Therefore, low or inconsistent correlations between component scores for inferencing and other variables do not provide much evidence against the hypothesis that these processes are particularly important.

A second line of evidence on the centrality of particular component processes comes from demonstrations that certain types of task manipulations are more likely than others to increase the Gf loading of a task (Pellegrino, 1985; Sternberg, 1986). Sternberg (1986) calls these selective

encoding, i.e., the requirement to attend selectively to information and to encode only that subset that is likely to be needed for solving a problem; selective comparison, i.e., to retrieve only information that is relevant to a problem, especially when the set of potentially relevant information in memory is vast; and selective combination, i.e., to assemble in working memory information already selected as relevant. Selective encoding depends heavily on the individual's store of prior knowledge (schema) and its attunement to the affordances of the situation. It also means the ability to resist the distractions of salient but irrelevant information, or, when solving items on mental tests, looking ahead to the alternatives before studying the stem (Bethell-Fox et al., 1984). Selective comparison also depends heavily on the store of knowledge, but also on its organization and accessibility, especially the ability to search rapidly through memory for overlap between two concepts. This is the essential feature of inference or abstraction problems: finding ways in which concepts A and B are not merely associated with each other, but rather finding the rules or relations that most specifically characterize their association. Problems in inductive reasoning emphasize selective encoding and comparison. Problems in deductive reasoning, on the other hand, emphasize selective combination. For example, syllogistic reasoning problems are difficult not because it is difficult to discern the relevant information in statements such as "all A are B" or in the understanding of the relations between words such as "all" and "some" (although this is a source of confusion for some); rather, the main difficulty is in keeping track of all of the ways in which the premises can be combined. This taxes both working memory and the ability to manipulate symbols. Thus, although certain processes may be central to intelligent thinking, individual differences in those processes may be in part due to other system limitations--
such as working-memory resources.

Attention and Working-Memory Capacity

All information-processing models of memory and cognition posit the existence of a limited capacity short-term or working memory that functions not only as a central processor, but as a bottleneck in the system. Some see this in terms of structure or capacity limitations; others view it in terms of attentional resources. Hunt and Lansman (1982) and Ackerman (1988)

argue that tasks that show higher correlations with G require more attentional resources. Attempts to manipulate the attentional demands of tasks often use a dual-task paradigm. Here, participants are required to do two things simultaneously, such as searching for a particular stimulus in a visual display while simultaneously listening for a specified auditory stimulus. Although the effect is often not observed, differences between more and less able subjects are typically greater in the dual task than in the single task condition. However, interpretation of this finding is problematic. For example, in one study, Stankov (1988) found that correlations with both Gc and Gf, but especially Gf, were higher for dual tasks than for single tasks. However, high levels of performance in the dual task situation were due to a strategy of momentarily ignoring one task while attending to the other. Thus, what on the surface seemed to implicate attentional resources on closer inspection implicated self-monitoring and the shifting of attentional resources.

Attentional requirements of tasks vary according to an individual's familiarity with the task and to the susceptibility of the task to automatization. Tasks--or task components--in which there is a consistent mapping between stimulus and response can be automatized in this way. Individuals who recognize the consistencies thus automatize task components more rapidly than those who are not so attuned. Thus, explanations of ability differences in terms of differences in attentional resources must account not only for attention shifting in dual tasks, but also differences in the extent to which task steps are or become automatized.

The explanation of differences in reasoning as reflecting differences in working memory capacity parallels the attentional explanation. Many researchers have claimed that a major source

of individual differences on reasoning tasks lies in how much information one must maintain in working memory, especially while effecting some transformation of that information (Holzman, Pellegrino, & Glaser, 1980). For example, as Kyllonen and Christal (1990) noted, most of the performance processes (such as encoding and inference) and executive processes (such as goal setting, goal management, and monitoring) are presumed to occur in working memory. Thus,

even though, say, the inference process may be effective, it must be performed within the limits of the working memory system. Therefore, although many different processes may be executed in the solution of a task, individual differences in them may primarily reflect individual differences in working memory resources.

Newer theories of working memory also differ importantly from the older concept of short-term memory. Indeed, one of the major differences lies in the relative emphasis on passive storage functions in older theories of short-term memory versus more effortful, controlled processing of information that is also being maintained in an active state in newer theories of working memory. Some see this in terms of a tradeoff between processing capacity and storage capacity (Daneman & Carpenter, 1980), whereas others view it in terms of different memory systems. For example, Baddeley (1986) posits a working memory with a passive storage component and a separate executive (or supervisory attentional system) that attends selectively to one stimulus while inhibiting another, coordinates performance in tasks, and switches strategies (Baddeley, 1996). When working memory is interpreted in this way, studies that find high correlation between working memory and reasoning seem less astonishing. For example, Kyllonen and Christal (1990) found that latent variables for reasoning ability and working memory correlated approximately $r = .8$ in four large studies. Their working memory tasks were specifically designed to reflect Baddeley's theory, and thus required both storage and transformation (although the former was presumed to be more difficult). Critics of the Kyllonen-Christal studies have argued that some of the tasks they used to measure working memory mirror common reasoning tasks. The criticism is not without merit. The same task was used as a working-memory test in one study and as a reasoning test in another study.

Oberauer, Süβ, Schulze, Wilhelm, and Wittman (1996) sought to address this problem. They administered 26 tasks designed to measure one or more of the three putative functions of working memory (storage and processing, monitoring, or coordination) on three contents (verbal, numerical, or figural) and the Berlin Intelligence test to a sample of 113 young adults. The Berlin Intelligence Scale is a faceted test that crosses operation (speed, memory, creativity, and

reasoning) with content (verbal, figural, numeric). Scores on both the intelligence test and the working-memory battery were aggregated in different ways (for example, by content or by operation) and to different levels (for example, to the level of test, the level of content or operation, or to the level of a single score). Analyses then focused on the utility of different parsings in accounting for relations between the working-memory tests and the intelligence test. Results showed that at the highest level of aggregation, latent factors from the working-memory battery and the intelligence test were highly related (estimated disattenuated correlation was approximately r = .92). However, differentiations of working memory into two functions (a first function of storage, processing, and coordination; a second function of supervision) and the test battery into subtests grouped by operation (speed, memory, creativity, and reasoning) led to a significant improvement in model fit. Other analyses showed that although the division between verbal/numerical content and spatial content was useful for the working-memory measures, the hypothesized three-way split into verbal, numerical, and spatial contexts was not.

In short, although the Oberauer et al. (1996) study supports the Kyllonen and Christal (1990) claim that individual differences in working memory are largely redundant with individual differences in reasoning ability, the study also suggests that working memory may usefully be broken down into at least two subprocesses (storage/processing and supervision) that operate with different effectiveness on two types of content (verbal/numerical and spatial). The argument is reminiscent of the earlier debate between Spearman and his American critics (first Thorndike, then Thurstone) as to whether the central processes of intelligence (which he identified with g) were unitary or multiple. And, as in the earlier debate, the argument seems to hinge upon the relative value of psychological meaningfulness of constructs (which usually favors decomposition) versus predictive ability (which usually favors aggregation).

Adaptive Processing

While acknowledging that individual differences in G reflected differences in all of these levels--in the speed and efficacy of elementary processes, in attentional or working memory resources, in the action of processes responsible for inference and abstraction (which includes

knowledge, skill, and attunement to affordances in the task situation)--several theorists have argued that more is needed. Sternberg (1985) argues that intelligent action requires the application of metacomponents--i.e., control processes that decide what the problem is, select lower-order components and organize them into a strategy, select a mode for representing or organizing information, allocate attentional resources, monitor the solution process, and attend to external feedback. Marshalek et al. (1983) emphasized the importance of assembly and control processes. They hypothesized 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, within as well as between items in a task, and between tasks" (Marshalek et al., 1983, p. 124). The Carpenter et al. (1990) analysis of the Raven test (see pp. ___) supports and extends this hypothesis. In their simulation, the crucial executive functions were (a) the ability to decompose
a complex problem into simpler problems and (b) the ability to manage the hierarchy of goals and subgoals generated by this decomposition.

The claim is <u>not</u> that more able problem solvers are always more strategic or flexible or reflective in their problem solving (cf. Alderton & Larson, 1994). Indeed, as previously noted in the discussion of strategies and strategy shifting, subjects who are most able to solve items often show little evidence of strategy shifting. For example, in the Kyllonen, Lohman, and Woltz (1984) study of a spatial synthesis task, subjects very high in spatial ability (but low in verbal ability) were best described by a model that said that they always mentally synthesized stimuli. These subjects probably did not have to resort to other strategies. Rather, it was the subjects who had less extreme scores profiles but relatively high scores on <u>G</u> that showed the most strategy shifting. The authors' interpretation was that, for this latter group of subjects, the pure "spatial" strategy of mentally combining figures exceeded working memory resources as problem difficulty increased. Subjects then switched to other strategies, as necessary, to solve problems. This required the ability to monitor one's problem solving as it proceeded, to assemble a new strategy when

necessary, and to make these adaptations as item demands varied. Indeed, the models they tested could be fit only to the extent that subjects systematically altered their solution strategies in response to observable features of items.

One can provide a stronger test of the hypothesis by turning the problem around. In other words, if fluid reasoning requires flexible adaptation, then it should be possible to manipulate the extent to which items require such processing and thereby alter their relationship with G. This was the approach taken by Swiney (1985) and by Chastain (1992). Swiney sought to test the hypothesis that correlations between performance on geometric analogies and G would increase as more flexible adaptation was required, at least for easy and moderately difficult problems. Correlations with G were expected to decline if task difficulty was too great. Adaptation was manipulated by grouping items in different ways. In the blocked condition, inter-item variation was minimized by grouping items with similar processing requirements (estimated by the number of elements, and the number and type of transformations). In the mixed condition, items were grouped to be as dissimilar as possible.

Two experiments were conducted. In the first study, 20 subjects, selected to represent the full range of ability in a pool of 146 high school students, were administered geometric analogy items that varied in difficulty (low versus high), condition (blocked or mixed), phase (rule learning, rule identification, rule application), and order (blocked or mixed first). In the rule learning phase, participants solved a series of five geometric analogies and verbalized the rule common to the set. Those who did not identify the rules were taught them. In the identification phase, subjects were required to determine which analogy stem differed from the others in a set of four stems. Rules

used to construct the stems had been taught in the learning phase. The application phase was identical to the rule identification phase except that subjects were given the correct rule(s) prior to the presentation of the item.

In the second experiment, 50 subjects representing the full range of G in another pool of high school students performed the same task, with the addition of a rule discovery phase and

items from sets C, D, and E of the Raven progressive matrices. In the discovery phase, subjects were first given practice on rules not used in the experiment, and then attempted 18 items. Raven items were split into 18 odd- and 18 even-numbered items. Items in the blocked condition were administered in this order; items in the mixed condition were reordered by interspersing the hardest items throughout the 18-item test.

Results of the first experiment showed that low-ability subjects were more adversely affected by mixing items than were high-ability subjects, regardless of treatment order. Effects were similar for the Raven in Experiment 2, but smaller. Relationships between task accuracy and

G for the different item sets in Experiment 2 is shown in Figure 10. Clearly, relationship with G varies systematically as a function of item difficulty and task requirements. Strongest relationships were observed for identifying (i.e., inferring) and applying difficult rules. Weakest relationships were observed for applying easy rules or discovering difficult rules, especially in the mixed condition.

Insert Figure 10 about here

Retrospective reports supported the conclusion that high-G subjects were better able to adapt their strategies flexibly to meet changing task demands. Low-G subjects reported a preference for holistic strategies such as trying to "see" the answer; high-G subjects reported switching to more analytic strategies as item difficulty increased. In contrast, low-G subjects were more likely to

report just "trying harder" on more difficult problems. Swiney also found that low-G subjects overestimated their performance on highly difficult items; they also consistently underestimated the difficulty of problems. This suggests differences in monitoring and evaluation processes.

Chastain (1992) reported three additional studies contrasting blocked versus mixed item presentations. Experiments 1 and 2 used items from the Wonderlic Personnel Test, a 50-item test that samples a broad range of item formats. The third experiment used a figural encoding task and a dynamic spatial task. In all studies, flexible adaptation was estimated by a simple difference score (mixed minus blocked) and by a residual score (regression of mixed on blocked). Correlations between these two scores, reference tests, and performance on a logic-gates learning task were small, but generally in the expected direction.

Gustafsson (in press) reports a study by Carlstedt that challenges this interpretation of the blocked-mixed contrast. Carlstedt administered three kinds of inductive reasoning problems to groups of Swedish military recruits: figure series completion (series), figure classification (groups) and a task called opposite groups (identify the feature that unites the figures in one group and differentiates the group from the other group). Items were combined in different ways to form

two test forms: heterogeneous (HET) and homogeneous (HOM). In the HET forms, subjects first

attempted one groups item, then one opposite groups item, and then one series item, after which the sequence was repeated. In the HOM form, groups item were presented first, then all series items, and then all opposite groups items.

Unexpectedly, Carlstedt found that G loadings were higher in the HOM condition than in the HET condition. He argues that the homogeneous arrangement affords better possibilities for learning and transfer across items. Different principles are introduced successively and combined in the more complex items. Gustafsson (in press) claims that the efficiency of a test as a measure of G is thus partly a function of dependence among the items. Such dependencies violate a basic assumption of IRT methods for scaling responses, once again illustrating the conflict between psychological and psychometric theories of test performance (Snow & Lohman, 1989).

To summarize the discussion to this point: as one moves from periphery to center in a two (or even three) dimensional radex, tasks increase in apparent complexity. Tasks near the center typically require more steps or component processes, and emphasize accuracy rather than speed

of response. But this does not mean that speed of processing is unimportant or that the addition of any type of process will increase the correlation with G. Increasing the demand on certain types processing, which Sternberg describes as selective encoding, comparison, and combination, also increases the correlation with G. Importantly, though, such processes require controlled, effortful processing and place heavy demands on working memory resources. They also require subjects to be more strategic or flexible or adaptive in their problem solving, or to learn rules from easy item that will be needed in combination to solve hard items.

Limitations and Future Directions

The information-processing paradigm has enormously enriched our understanding of cognitive tests and the ability constructs they estimate. We have moved from trait labels and vague notions of "process" to the rich and detailed models of thinking sampled in this chapter. All paradigms are inadequate in some respects, and the information-processing approach is no

exception. Two shortcomings are particularly salient: (a) the neglect of affect and conation, and (b) the failure to understand the contextual specificity of abilities.

Affect and conation. Although theorizing about the influence of affect (or feeling) and conation (or willing) on cognition dates back to the Greek philosophers, it is only recently that investigators have attempted to study their complex and reciprocal influences on each other. Although many promising leads have been identified (see Snow, Corno, & Jackson, 1996; Boekaerts, 1995; Kuhl & Kraska, 1989; Corno & Kanfer, 1993; Schunk & Zimmerman, 1994), there is no simple way to summarize the enormous diversity of paradigms and findings. Nonetheless, it is clear that persons who do well on ability tests expend effort differently from persons who score poorly. In general, those who score well retrieve information from memory with greater ease and rapidity, and are better able to maintain that information in working memory while concurrently executing other processes. The difference is most striking in comparisons of experts and novices in skill domains such as reading. Experts expend their efforts on high-level processes (that include, but go beyond comprehension), whereas novices struggle to identify words and the sentences they create. Affect enters not only as anxiety and frustration, which further constrict cognition, but also as interest and surprise, which enhance and direct cognition. In particular, those who adopt a constructive motivational orientation towards a task will tend to exhibit more and better self-regulation than individuals who adopt a less-constructive or even defensive orientation. Situations differentially elicit these conative and affective resources. Indeed, understanding the role of affect in cognition seems to demand a mode of theorizing and experimentation that attends not only to persons or to situations, but also to the attunement of particular individuals to particular aspects of situations.

Including situations and their affordances. A theory of G must explain individual differences in problem solving not only on tests, but in school and other everyday contexts. Although occasionally nodding to the role of culture, cognitive theories of abilities have not taken seriously the fact that cognition is situated. Snow (1994) argues that a theory of G needs to start with the proposition that abilities are situated. In his view, this means that abilities are

reflected in the tuning of particular persons to the particular demands and opportunities of situations, and thus reside in the union of person in situation, not "in the mind" alone.

> The situation contains some pieces of what the person needs or can use to accomplish a given task. But persons must be tuned to perceive and use these pieces, and also to supply needed pieces from their own learning histories. Some persons are prepared to perceive these affordances, to use the pieces provided by the situation, and to complement these with pieces they provide, but some are not. Among those who are so tuned, each may use and supply slightly different pieces; there is functional equivalence despite idiosyncrasy. The result is that some persons succeed in learning in a given situation; they are in harmony with it. Others do not, because they are not tuned to use the opportunities the situation provides or to produce what it demands. (Snow, 1994, p. ___).

The idea of affordances brings back not only the physical and social environment, but also the particular couplings or attunements to aspects of that environment that arise through the long history of the evolution of the species or the short history of the development of the individual. Put differently, the notion of affordances in the situation and propensities in the individual provides one way to reason about the selectivity in encoding exhibited by more able individuals (cf. Sternberg, 1986).

A summary hypothesis. Elshout (1985) defines a problem as the state of a particular person in a particular task situation such that, should the person succeed, any explanation of the event based only on the person's experience in that particular situation is excluded beforehand. Tasks that can be accomplished using stored routines do not, in Elshout's view, count as problems. Sternberg (1986) makes a similar but less radical distinction when he claims that processes used to solve problems must be executed in a controlled, rather than an automatized, fashion if reasoning is required. Likewise, Belmont and Mitchell (1987) contend that learners will generally be most strategic on tasks they perceive to be of moderate difficulty. This is because strategy use is

presumed to require effort and people are unlikely to invest effort when it is unnecessary or unlikely to be unrewarded.

Snow (1989) offers the following summary hypothesis. Imagine tasks as scaled along a continuum of difficulty or complexity. Elshout argues that there is a threshold for each person along such a continuum. Below the threshold, performance follows directly from routines the person already has in store for the task; the flow of activity is relatively automatic and algorithmic. Errors come mostly from cognitive slips and inattention rather than from inadequacies in the system. In the language of Cattell (1963), these are crystallized abilities and skills. Above

the threshold, however, the person must increasingly operate in a heuristic, improvisational, controlled, and achievement-motivated mode. Here, errors occur because the previously stored routines and knowledge are inadequate, or poorly applied, or are not tuned to the specific task at hand, or because motivation flags prematurely. Furthermore, the farther above one's threshold one is forced to work, the more likely that heuristic processing and improvising degrade into helpless, even anxious, muddling; errors become more conceptual and strategic. What have been here called fluid reasoning abilities would be measured for most individuals in the lower end of this range. Novices are thus to be seen as persons who must work at or above their thresholds in tasks of a given type, whereas experts are those who can work well below their thresholds in that task type. The contrast is also seen in the pattern of declining correlations with G and increasing correlations with more specific abilities as participants acquire a new skill (Ackerman, 1986, 1988). The goal of instruction is to move a person's threshold up, in each type of task that society and the person values. Raising the threshold means making more and more difficult or complex instances of the task type nonproblematical and automatic. To measure reasoning, however, problems must be perceived by the individual as falling somewhere above the lower threshold and below the upper threshold, i.e., within the zone of tolerable problematicity.

Conclusions and Implications

Desiderata for a Theory of Complex Information Processing

Any theory of intelligence that purports to explain the sort of complex information processing discussed in this chapter must accommodate several facts. First, the theory must explain the repeatedly demonstrated finding that human abilities are organized hierarchically (Carroll, 1993; Gustafsson & Undheim, 1996). Thus, theories that posit only a series of abilities arranged nonhierarchically (e.g., Thurstone, 1938; Gardner, 1983) are fundamentally inadequate. Second, the theory must explain the clustering of abilities by content, especially verbal, spatial, and numerical/symbolic. Thus, theories that posit only one individual-difference dimension (such as g) or claim that intelligence is reflected in the action of a single cognitive structure or process (such as attention or working memory) are also inadequate. Third, the theory must give some principled account of processing complexity and how it is related to the loading of a task on the general factor. Thus, theories that do not distinguish among levels of processing complexity are inadequate. Fourth and perhaps most importantly, the theory must coordinate the findings of differential psychology with general theories of human information processing. Theories of cognition that ignore the literature on individual differences are unlikely to provide a full accounting of the individual-difference construct of intelligence. Conversely, theories of individual differences that ignore the literature on human cognition are unlikely to evolve measures that are psychologically transparent, or that represent in a systematic way different features of the human cognitive system.

Theory-Based Tests and Testing

This chapter shows how information-processing psychology can be applied to existing ability tests and constructs in order better to understand them. But the problem can be turned around, and information-processing theories of cognition can be used to guide the development of new tests. Although Sternberg's (1985) triarchic theory of intelligence is in part rooted in the information-processing paradigm, an operational test that embodies all aspects of the triarchic theory has not been developed. Indeed, the main implication for assessment seems to be in the measurement of practical intelligence rather than in refined measures of fluid and crystallized abilities. The most extensive effort to date to develop an ability testing battery grounded in information-processing psychology has been the Cognitive Abilities Measurement (CAM)

program in the United States Air Force (Kyllonen, 1993, 1994). The CAM framework posits a 4 ∞ 3 grid with four "sources" and three "contents." The sources are processing speed, working memory, declarative knowledge, and procedural knowledge. The three contents are verbal, quantitative, and spatial. Empirical tests of the model confirm that both content and source (or process) can be distinguished, and also show that a hierarchical model fits the data slightly better than a flat or nonhierarchical model (Kyllonen, 1993).

The CAM framework is tied to theories of cognition in two ways: (1) through the structure of the framework itself, particularly the four sources, and (2) through the tasks used to represent different cells in the model. For example, working-memory tasks were based on Baddeley's (1986) theory of working memory, and in some cases, directly modeled after tasks used in his research program or that of other researchers (e.g., Daneman & Carpenter, 1980). It is much easier to elaborate ties to cognitive processes and to address the construct validity of tasks when theory is used in this way to guide test construction rather than post hoc to guide interpretation of investigations of otherwise ambiguous tasks. On the other hand, tests for some cells in the model were not constructed, and tests for other cells were selected on the basis of availability or apparent match to the somewhat vague source definitions. Further work is needed to establish the validity of task selection and assignment to cells in the framework--perhaps, for example, by panels of judges. The CAM battery is also interesting for what it tells us about how best to measure processes. Early in the project, much effort was devoted to examining the correlates of various "process" measures estimated by individual regression coefficients or simple difference scores. None of these measures were retained in the final battery. Instead, individual performance is summarized in various total (or part) scores on tasks designed to emphasize a particular aspect of individual differences in processing. Once again, it seems that information-processing models--while extremely useful for the internal validation of tasks--are not particularly helpful when it comes to the practical task of specifying dependable individual scores with stable external correlates.

The CAM battery uses theory at a fairly global level to guide task selection. Other efforts

to use theory to guide test construction have generally used process models of particular tasks to guide the development of items for new tests (see, e.g., Embretson, in press; Bejar, 1985; Irving, Dann, & Anderson, 1990; Nichols, 1994). Theory-based tests offer many advantages, including the possibility of constructing items with predictable characteristics "on the fly," an especially attractive option in computer adaptive testing. There are also advantages for creation of paper-and-pencil tests. For example, estimating the difficulty of an item without first administering it can eliminate an entire cycle in test development. There are drawbacks, however. Verbal reasoning items have proven vastly more difficult to generate in this way. The substantial effort of Bejar et al. (1991) to generate verbal analogy items for the SAT did not succeed. The more recent efforts of Buck et al. (1998) using Tatsuoka's (1995) methods may be more successful, at least in identifying more of the many item attributes that influence performance on verbal analogies. A less-obvious difficulty with theory-based tests lies in the relatively constricted set of items that are generated by a particular set of rules. The hodgepodge of items on many well-constructed ability tests may actually require more flexibility from test takers and may also be less amenable to coaching or simple practice effects than pools of items generated by more restricted sets of rules. These concerns apply primarily to Gf measures rather than to measures of more-specialized abilities such as perceptual speed or even working-memory capacity.

Sources of Difficulty, Sources of Individual Differences

Understanding what makes a task difficult is not the same as understanding how participants solve items on the task, but it is a useful place to start. Indeed, although there are often many different ways to solve a task that would conform with the observation that some types of items are more difficult than others, there is a much larger set of processing models that are excluded by an observed ordering of items according to difficulty. More important, though, is the realization that understanding how participants solve items on a task is not an understanding of individual differences--except in the rare instance in which individual differences are entirely reflected in how the task is solved. Manipulations that make tasks difficult will, in general, make people differ. But making a task difficult is like making an object

heavy--there are many, many ways to make something heavy. Some of these manipulations introduce construct relevant variance (making a bucket of water heavier by adding more water); some introduce construct irrelevant variance (making the bucket out of lead). An analysis that partitions variation into stimulus (or task) variance and individual difference (or subject) variance can help clarify these relationships (see Embretson, 1985; this volume). Indeed, one of the major contributions of cognitive-process analyses of ability tests to the practical business of measurement has been to identify irrelevant sources of difficulty in tasks and conditions in which they are eliminated (Embretson, 1985; Snow & Lohman, 1989). However, even individual difference variance

may be construct irrelevant, as was shown the examples (see pp. ____) where increases in task difficulty beyond some point caused a decrease in correlation with the target construct, and an increase in correlation with some other construct. The main point, however, is that isolated analyses of particular tasks--no matter how well selected--cannot support a theory of an individual difference construct. For this reason, Sternberg and his colleagues, Pellegrino and his colleagues, and Snow and his colleagues all sought to study families of related tasks. There seems to be no other way to discover which portions of the individual-difference variance are task-specific and which portions generalize to other tasks.

We now also know that understanding individual differences on tasks used to measure general reasoning abilities requires the investigation of items that are not only complex in the sense of requiring multiple processes, but also difficult. In other words, conclusions about process differences between high- and low-ability subjects depend importantly on the difficulty of the task. In particular, recent studies of individual differences in inductive reasoning replicate earlier findings that high-ability subjects spend more time encoding and less time on inference and application, and are more likely to engage in exhaustive processing. However, these studies also show quite different, even contrary, patterns on difficult problems.

Items contain options, not just stems. How subjects use information in options to reason about the problem posed in the stem is an important aspect of how they solve difficult reasoning

items, especially classification problems. Younger and less-able subjects are often misled or distracted by foils. Older and more-able subjects show evidence of both working backward and working forward on such problems. What might appropriately be described as "justification" or cycling through the stem a second time on simple problems requires a far more extensive set of processes on complex problems. Complex problems also require a level of task decomposition, goal setting, and monitoring unlike that required on simple problems (Carpenter et al., 1990). Once again, early suspicions that

complex problem solving required a level of planning (or "assembly") and monitoring (or "control") processes have been confirmed. Now some argue that there are affective and volitional states and processes to be considered as well (Snow et al., 1996).

On the other hand, we have learned that the lion's share of individual differences in inductive reasoning is shared with working memory. This is an extremely important finding. Importantly, though, working memory is represented by tasks that require a good deal of higher-order or executive processes (Baddeley, 1996).

Do Cognitive Tests Make Good Cognitive Tasks?

Most of the research reviewed in this chapter has followed Estes' (1974) argument that the best way to understand intelligence is to understand how individuals solve items on intelligence and other ability tests. This was certainly a reasonable place to start. But ability tests are limited in at least two ways. First, some argue that the abilities sampled by such tests are an unrepresentative sample of the full range of human competencies that properly fall under the rubric of intelligent behavior. These critics argue that measures of musical abilities, social intelligence, practical intelligence, everyday reasoning, participatory skills, etc., should be studied as well. As long

as we have no agreed upon definition of the universe of competencies that qualify as "intelligent," then there will always be reasonable (as well as unreasonable) criticisms of this sort.

Ability tests are limited in another respect, however. Indeed, even if such tests were thought to constitute a fair and representative sample of the behavior considered intelligent, one might still argue that psychometric tests do not constitute the best or most revealing way to study the processes which generate such behaviors. The claim here is that although psychometric tests often constitute efficient ways to measure individual differences in abilities, they often are not very informative vehicles for understanding what those abilities might be. For example, although a multiple-choice vocabulary test is an excellent measure of both verbal ability and general ability, even the most careful information-processing analysis of how subjects respond to items on such tests reveals little about the nature of either verbal or general ability. Because of

this, some have sought instead to understand abilities by correlating ability test scores with scores derived from laboratory tasks (e.g., Hunt, Frost, & Lunneborg, 1973) or from learning tasks (e.g., Ackerman, 1988). Process models of any of these tasks--whether based on ability tests or laboratory tasks, or learning tasks--are often most informative when subjects who differ in ability differ systematically in how they process information. Such qualitative differences have been pivotal in investigations of cognitive development (e.g., Piaget, 1963). On this view, then, a process understanding of ability requires the invention or selection of tasks that elicit qualitative differences in processing between individuals who differ in ability. Scores on cognitive tests provide an important external reference point, but may not be themselves the most informative objects of inquiry into the nature of abilities. This is particularly the case for ability tests and constructs that primarily reflect the efficacy of past processing (e.g., a test of verbal knowledge) rather than of present processing (e.g., a test of figural analogical reasoning).

Methodology

Methodology matters. The seemingly straightforward task of estimating process scores for individuals by subtracting latency in one condition from latencies in another turned out to be far more troublesome than initially envisioned. What should be done about error-response latencies? The experimenter not interested in individual differences often eliminates subjects who err too often, and does not worry if all remaining participants are equated on speed-accuracy tradeoff. Not so when individual differences are the object of investigation. Further, on any but the simplest RT task, multiple processes must be used. Thus, subtraction gave way to more complex regression procedures, which estimated scores for several components simultaneously and which also gave some indication of overall model fit. Attempts to account for both error and latency led some to use canonical correlation, and others to control exposure latencies experimentally and to fit nonlinear regression models. Others sought to improve the scaling of accuracy and latency data through transformations, signal-detection theory, and latent-trait models (as in Embretson's multicomponent latent-trait models). Tatsuoka's rule space procedure (see Figure __) shows the current state of one effort to solve some of these problems.

Thus, methodology matters, perhaps as much for how every attempt to solve one problem has created others. Both researchers and outsiders have been led down more than one blind alley by the failure to understand the sometimes not-so-obvious limitations of their favorite methodology.

What does the future hold? If the recent past is any guide, we will see the further development of sophisticated methodologies for modeling performance on complex tasks. But it is unlikely that such procedures will be widely used. If Sternberg's (1977) componential methods--which were based on relatively straightforward regression procedures--were too daunting for many, how will multicomponent latent-trait theory or the rule space methodology fare? On the bright side, confirmatory factor analyses and modeling techniques have shown structure where chaos once ruled (see especially Gustafsson, 1988); new scaling procedures in psychometrics allow comparisons of items across samples (Embretson, 1985; Sheehan, 1997), and the development of intelligent systems for creating items and scoring complex constructed responses (Bennett, 1998). These and other developments will continue to inform and reform assessment of complex information processing. Thus, the future will bring continued innovations, but few that are simple. Whether the constructs
of the future will look radically different from those of the past is more difficult to say. In spite of repeated arguments against G and in favor of one or another system of multiple abilities, data routinely support not so much the existance of G as its utility. Thus, it is a fair bet that, whatever the future holds, tests of G--especially Gf and particularly I--will be a part of it.

References

Ackerman, P. L. (1986). Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. Intelligence, 10, 101-139.

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: A theory of cognitive abilities and information processing. Journal of Experimental Psychology: General, 117, 299-329.

Ackerman, P. L., & Kanfer, R. (1993). Integrating laboratory and field study for improving selection: Development of a battery for predicting air traffic controller success. Journal of Applied Psychology, 78, 413-432.

Alderton, D. L., & Larson, G. E. (1994). Cross-task consistency in strategy use and the relationship with intelligence. Intelligence, 18, 47-76.

Alderton, D. L., Goldman, S. R., & Pellegrino, J. W. (1985). Individual differences in process outcomes for verbal analogy and classification solution. Intelligence, 9, 69-85.

Anastasa, A., & Foley, J. P. (1949). Differential psychology (rev. ed.). New York: Macmillan.

Anderson, G. V., Fruchter, B., Manuel, H. T., & Worchel, P. (1954). Survey of research on spatial factors (Research Bulletin AFPTRC-TR-54-84). San Antonio, TX: Lackland AFB.

Anderson, J. R. (1976). Language, memory, and thought. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1985). Cognitive psychology and its implications (2nd ed.). New York: W. H. Freeman.

Anderson, J. R. (1993). Rules of the mind. Hillsdale, NJ: Erlbaum.

Baddeley, A. (1996). Exploring the central executive. The Quarterly Journal of Experimental Psychology, 49A(1), 5-28.

Baddeley, A. D. (1986). Working memory. Oxford: Clarendon Press.

Bejar, I. I. (1985). Speculations on the future of test design. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 279-294). New York: Academic Press.

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). Cognitive and psychometric analysis of analogical problem solving. New York: Springer-Verlag.

Belmont, J. M., & Butterfield, E. C. (1971). Learning strategies as determinants of memory deficiencies. Cognitive Psychology, 2, 411-20.

Belmont, J. M., & Mitchell, D. W. (1987). The general strategy hypothesis as applied to cognitive theory in mental retardation. Intelligence, 11, 91-105.

Bennett, R. E. (1998). Reinventing assessment: A policy information perspective (A report from Speculations on the future of large-scale educational testing). Princeton, NJ: Educational Testing Service, Policy Information Center, Research Division.

Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity, familiarity, and individual differences. Journal of Experimental Psychology: Human Perception and Performance, 14, 12-23.

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

Biggs, J. B., & Collis, K. F. (1982). Evaluating the quality of learning: The SOLO Taxonomy. New York: Academic Press.

Binet, A., & Henri, V. (1896). La psychologie individuelle [Individual psychology]. Année psychologique, 11, 191-465.

Bloom, B. S., & Broder, L. J. (1950). Problem-solving processes of college students. Supplementary Educational Monographs, 73.

Boekaerts, M. (1995). The interface between intelligence and personality as determinants of classroom learning. In D. H. Saklofske & M. Zeidner (Eds.), International handbook of personality and intelligence (pp. 161-183). New York: Plenum Press.

Brody, N. (1992). Intelligence (2nd ed.). San Diego, CA: Academic Press.

Buck, G., VanEssen, T., Tatsuoka, K., & Kostin, I. (1998). The process of identifying cognitive and linguistic attributes underlying performance in the verbal domain: An example with the SAT-V. Unpublished paper. Princeton, NJ: Educational Testing Service.

Butterfield, E. C., Nielsen, D., Tangen, K. L., & Richardson, M. B. (1985). Theoretically based psychometric measures of inductive reasoning. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 77-148). New York: Academic Press.

Carpenter, P. A., & Just, M. A. (1978). Eye fixations during mental rotation. In J. W. Senders, D. F. Fisher, & R. A. Monty (Eds.). Eye movements and the higher psychological functions. Hillsdale, NJ: Erlbaum.

Carpenter, P. A., Just, M. A., & Schell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. Psychological Review, 97, 404-431.

Carroll, J. B. (1941). A factor analysis of verbal abilities. Psychometrika, 6, 279-307.

Carroll, J. B. (1993). Human cognitive abilities. A survey of factor-analytic studies. Cambridge, UK: Cambridge University Press.

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

Chaffin, R., & Herrmann, D. J. (1984). The similarity and diversity of semantic relations. Memory & Cognition, 12, 134-141.

Chastain, R. L. (1992). Adaptive processing in complex learning and cognitive performance. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Christal, R. E. (1958). Factor analytic study of visual memory. Psychological Monographs, 72 (13: Whole No. 466).

Clark, H. H. (1969). The influence of language in solving three-term series problems. Journal of Experimental Psychology, 82, 205-215.

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

Cooper, L. A. (1982). Strategies for visual comparison and representation: Individual differences. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 1, pp. 77-124). Hillsdale, NJ: Erlbaum.

Cooper, L. A., & Regan, D. T. (1982). Attention, perception, and intelligence. In R. J. Sternberg (Ed.), Handbook of human intelligence (pp. 123-169). Cambridge, UK: Cambridge University Press.

Corno, L, & Kanfer, R. (1993). The role of volition in learning and performance. In L. Darling-Hammond (Ed.), Review of research in education (Vol. 19, pp. 3-43). Washington, DC: American Educational Research Association.

Crawford, J. (1988). Intelligence, task complexity and tests of sustained attention. Unpublished doctoral dissertation, University of New South Wales, Sydney, Australia.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. American Psychologist, 12, 671-684.

Damasio, A. (1994). Descartes error: Emotion, reason, and the human brain. New York: Putnam.

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

Deary, I. J., & Stough, C. (1996). Intelligence and inspection time: Achievements, prospects, and problems. American Psychologist, 51, 599-608.

Detterman, D. K. (1986). Human intelligence is a complex system of separate processes. In R. J. Sternberg & D. K. Detterman (Eds.), What is intelligence? (pp. 57-61). Norwood, NJ: Ablex.

Diones, R., Bejar, I., & Chaffin, R. (1996). The dimensionality of responses to SAT analogy items (Research Report 96-1). Princeton, NJ: Educational Testing Service.

Ebbinghaus, H. (1985). Memory: A contribution to experimental psychology (H. A. Ruger & C. E. Bussenues, Trans.). New York: Teachers College Press. (Original work published 1913)

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). Kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Elshout, J. J. (1985, June). Problem solving and education. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179-197.

Embretson, S. E. (1985). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 195-218). New York: Academic Press.

Embretson, S. E. (1986). Intelligence and its measurement: Extending contemporary theory to existing tests. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 3, pp. 335-368). Hillsdale, NJ: Erlbaum.

Embretson, S. E. (1995). The role of working memory capacity and general control processes in intelligence. Intelligence, 20, 169-189.

Embretson, S. E., Schneider, L., & Roth, D. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. Journal of Educational Measurement, 23, 13-32.

Erickson, J. R. (1974). A set analysis theory of behavior in formal syllogistic reasoning tasks. In R. Solso (Ed.), Loyola symposium on cognition (Vol. 2, pp._____). Hillsdale, NJ: Erlbaum.

Erickson, J. R. (1978). Research on syllogistic reasoning. In R. Revlin & R. E. Mayer (Eds.), Human reasoning (pp._____). Washington, DC: Winston.

Ericsson, K. A., & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.

Estes, W. K. (1974). Learning theory and intelligence. American Psychologist, 29, 740-749.

Evans, T. G. (1968). A program for the solution of geometric-analogy intelligence test questions. In M. Minsky (Ed.), Semantic information processing. Cambridge, MA: MIT Press.

Frederiksen, J. R. (1981). Sources of process interaction in reading. In A. M. Lesgold & C. A. Perfetti (Eds.), Interactive processes in reading (pp. 361-386). Hillsdale, NJ: Erlbaum.

Gardner, H. (1983). Frames of mind: The theory of multiple intelligences. New York: Basic Books.

Gentile, J. R., Kessler, D. K., & Gentile, P. K. (1969). Process of solving analogy items. Journal of Educational Psychology, 60, 494-502.

Gitomer, D. H., Curtis, M. E., Glaser, R., & Lensky, D. B. (1987). Processing differences as a function of item difficulty in verbal analogy performance. Journal of Educational Psychology, 79, 212-219.

Goldman, S. R., & Pellegrino, J. W. (1984). Deductions about induction: Analyses of developmental and individual differences. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 2, pp. 149-197). Hillsdale, NJ: Erlbaum.

Goldman, S. R., Pellegrino, J. W., Parseghian, P. E., & Sallis, R. (1982). Developmental and individual differences in verbal analogical reasoning by children. Child Development, 53, 550-559.

Greeno, J. G. (1978). A study of problem solving. In R. Glaser (Ed.), Advances in instructional psychology (Vol. 1, pp. 13-75). Hillsdale, NJ: Erlbaum.

Greeno, J. G. (1980). Psychology of learning, 1960-1980: One participant's observations. American Psychologist, 35, 713-728.

Guilford, J. P., & Lacey, J. I. (Eds.). (1947). Printed classification tests. Army Air Forces aviation psychology research program (Report No. 5). Washington, DC: Government Printing Office.

Gustafsson, J-E. (in press). Measuring and understanding G: Experimental and correlational approaches. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), Learning and individual differences: Process, trait, and content determinants. Washington, DC: American Psychological Association.

Gustafsson, J.-E. (1988). Hierarchical models of individual differences in cognitive abilities. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 4, pp. 35-71). Hillsdale, NJ: Erlbaum.

Gustafsson, J.-E., & Undheim, J. O. (1996). Individual differences in cognitive functions. In D. C. Berliner & R. C. Calfee (Eds.), Handbook of educational psychology (pp. 186-242.) New York: Simon & Schuster Macmillan.

Hagendorf, H., & Sá, B. (1995). Koordinierungsleistungen im visuellen Arbeitsgedächtnis [Coordination in visual working memory]. Zeitschrift für Psychologie, 203, 53-72.

Halford, G. S., Maybery, M. T., O'Hare, A. W., & Grant, P. (1994). The development of memory and processing capacity. Child Development, 65, 1338-1356.

Heller, J. I. (1979). Cognitive processing in verbal analogy solution. (Doctoral dissertation, University of Pittsburgh), Dissertation Abstracts International, 40, 2553A.

Henley, N. M. (1969). A psychological study of the semantics of animal terms. Journal of Verbal Learning and Verbal Behavior, 8, 176-184.

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

Horn, J. L. (1972). The structure of intellect: Primary abilities. In R. M. Dreger (Ed.), Multivariate personality research. Baton Rouge, LA: Claitor.

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

Hunt, E. B., Frost, N., & Lunneborg, C. (1973). Individual differences in cognition: A new approach to intelligence. In G. Bower (Ed.), The psychology of learning and motivation: Vol. 7 (pp. 87-122). New York: Academic Press.

Hunt, E., & Lansman, M. (1982). Individual differences in attention. In R. J. Sternberg (Ed.), Advances in the psychology of human abilities (Vol. 1, pp. 207-254). Hillsdale, NJ: Erlbaum.

Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. Psychological Review, 75, 550-560.

Irvine, S. H., Dann, P. L., & Anderson, J. D. (1990). Towards a theory of algorithm-determined cognitive test construction. British Journal of Psychology, 81, 173-195.

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

Janser, R., De Boeck, P., & Vander Steene, G. (1996). Verbal fluency and verbal comprehension abilities in synonym tasks. Intelligence, 22, 291-310.

Jensen, A. R. (1980). Bias in mental testing. New York: Free Press.

Jensen, A. R. (1982). The chronometry of intelligence. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 1, pp. 255-310). Hillsdale, NJ: Erlbaum.

Jensen, A. R. (1987). Process differences and individual differences in some cognitive tasks. Intelligence, 11, 107-136.

Jensen, A. R. (1998). The g factor: The science of mental ability. Westport, CT: Praeger.

Johnson-Laird, P. N. (1972). The three-term series problem. Cognition, 1, 57-82.

Johnson-Laird, P. N. (1983). Mental models: Towards a cognitive science of language, inference and consciousness. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1985). Deductive reasoning ability. In R. J. Sternberg (Ed.), Human abilities: An information-processing approach (pp. 173-194). New York: W. H. Freeman.

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. Cognitive Psychology, 10, 64-99.

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. (1992). A capacity theory of comprehension: Individual differences in working memory. Psychological Review, 99, 122-149.

Kintsch, W., & Greeno, J. G., (1985). Understanding and solving word arithmetic problems. Psychological Review, 92, 109-129.

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

Kuhl, J., & Kraska, K. (1989). Self-regulation and metamotivation: Computational mechanisms, development, and assessment. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), Abilities, motivation, and methodology (pp. 343-374). Hillsdale, NJ: Erlbaum.

Kyllonen, P. C. (1984). Information processing analysis of spatial ability (Doctoral dissertation, Stanford University). Dissertation Abstracts International, 45, 819A.

Kyllonen, P. C. (1994). CAM: A theoretical framework for cognitive abilities measuremnt. In D. Detterman (Ed.), Current topics in human intelligence: Vol. 4. Theories of intelligence (pp. 307-359). Norwood, NJ: Ablex.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! Intelligence, 14, 389-433.

Kyllonen, P. C., Lohman, D. F., & Snow, R. E. (1984). Effects of aptitudes, strategy training, and task facets on spatial task performance. Journal of Educational Psychology, 76, 130-145.

Kyllonen, P. C., Lohman, D. F., & Woltz, D. J. (1984). Componential modeling of alternative strategies for performing spatial tasks. Journal of Educational Psychology, 76, 1325-1345.

Lansman, M. (1981). Ability factors and the speed of information processing. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), Intelligence and learning (pp. 441-457). New York: Plenum Press.

Law, D. J., Morrin, K. A., & Pellegrino, J. W. (1995). Training effects and working memory contributions to skill acquisition in a complex coordination task. Learning and Individual Differences, 7, 207-234.

LeFevre, J.-A., & Bisanz, J. (1986). A cognitive analysis of number-series problems: Sources of individual differences in performance. Memory & Cognition, 14, 287-298.

Lesgold, A. M. (1984). (1984). Acquiring expertise. In J. R. Anderson & S. M. Kosslyn (Eds.), Tutorials in learning and memory (pp. 31-60). New York: W. H. Freeman.

Lohman, D. F. (1979). Spatial ability: A review and reanalysis of the correlational literatures (Tech. Rep. No. 8). Stanford, CA: Stanford University, Aptitude Research Project, School of Education. (NTIS No. AD-A075 973).

Lohman, D. F. (1986). The effect of speed-accuracy tradeoff on sex differences in mental rotation. Perception and Psychophysics, 39, 427-436.

Lohman, D. F. (1988). Spatial abilities as traits, processes, and knowledge. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 4, pp. 181-248). Hillsdale, NJ: Erlbaum.

Lohman, D. F. (1994). Component scores as residual variation (or why the intercept correlates best). Intelligence, 19, 1-11.

Lohman, D. F., & Kyllonen, P. C. (1983). Individual differences in solution strategy on spatial tasks. In R. F. Dillon & R. R. Schmeck (Eds.), Individual differences in cognition (Vol. 1, pp. 105-135). New York: Academic Press.

Lohman, D. F., & Nichols, P. D. (1990). Training spatial abilities: Effects of practice on rotation and synthesis tasks. Learning and Individual Differences, 2, 69-95.

Lorge, I., & Thorndike, R. L. (1957). The Lorge-Thorndike intelligence test, levels A-H. Boston: Houghton-Mifflin.

Macleod, C. M., Hunt, E. B., & Mathews, N. N. (1978). Individual differences in the verification of sentence-picture relationships. Journal of Verbal Learning and Verbal Behavior, 17, 493-508.

Marquer, J., & Pereira, M. (1987, April). Individual differences in sentence-picture verification. Paper presented at the meeting of the American Educational Research Association, New York.

Marshalek, B. (1981). Trait and process aspects of vocabulary knowledge and verbal ability (Tech. Rep. No. 15). Stanford, CA: Stanford Univeristy, Aptitude Research Project, School of Education. (NTIS No. AD-A102 757).

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. Intelligence, 7, 107-128.

McGee, M. (1979). Human spatial abilities: Sources of sex differences. New York: Praeger.

Melis, C. (1997). Intelligence: A cognitive-energetic approach. Wageningen, The Netherlands: Ponsen & Looijen BV.

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

Mumaw, R. J., Pellegrino, J. Wl, Kail, R. V., & Carter, P. (1984). Different slopes for different folks: Process analysis of spatial aptitude. Memory & Cognition, 12, 515-521.

Newell, A. (1980). Resoning, problem-solving, and decision processes: The problem space as a fundamental category. In R. Nickerson (Ed.), Attention and performance VIII (pp. 693-718). Hillsdale, NJ: Erlbaum.

Newell, A., & Simon, H. A. (1961). GPS, a program that simulates human thought. In E. A. Feigenbaum & J. Feldman (Eds.), Computers and thought (pp. 279-96). New York: McGraw-Hill. (Reprinted from H. Billing [Ed.], Lernende Automaten. Munich: Oldenbourg.)

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

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessment. Review of Educational Research, 64, 575-603.

Oberauer, K. (1993). Die Koordination kognitiver Operationen – eine Studie zum Zusammenhang von Intelligenz und "working memory" [The coordination of cognitive operations – A study on the relation between intelligence and working memory]. Zeitschrift für Psychologie, 201, 57-84.

Oberauer, K., Süβ, H-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (1996). Working memory capacity - Facets of a cognitive ability construct. (Berichte des Lehrstuhls Psychologie II, Universität Mannheim [University of Mannheim Department of Psychology II Reports] Heft 7, Jahrgang 1996) Mannheim, Germany: University of Mannheim, Lehrstuhl für Psychologie II [Department of Psychology].

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. Cognitive Psychology, 9, 441-474.

Pellegrino, J. W. (1985). Inductive reasoning ability. In R. J. Sternberg (Ed.), Human abilities: An information-processing approach (pp. 195-225). New York: W. H. Freeman.

Pellegrino, J. W., & Glaser, R. (1980). Components of inductive reasoning. In R. E. Snow, P.-A. Federico, & W. E. Montague (Eds.), Aptitude, larning, and instruction: Vol. 1. Cognitive process analyses of aptitude (pp. 177-218). Hillsdale, NJ: Erlbaum.

Pellegrino, J. W., & Glaser, R. (1982). Analyzing aptitudes for learning: Inductive reasoning. In R. Glaser (Ed.), Advances in instructional psychology (Vol. 2, pp. 269-345). Hillsdale, NJ: Erlbaum.

Pellegrino, J. W., & Kail, R. (1982). Process analyses of spatial aptitude. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 1, pp. 311-366). Hillsdale, NJ: Erlbaum.

Pellegrino, J. W., & Lyon, D. R. (1979). The components of a componential analysis [Review of the book Intelligence, information processing and analogical reasoning: The compoenetial analysis of human abilities]. Intelligence, 3, 169-186.

Piaget, J. (1963). The psychology of intelligence. New York: International Universities Press.

Pieters, J. P. M. (1983). Sternberg's additive factor method and underlying psychological processes: Some theoretical considerations. Psychological Bulletin, 93, 411-426.

Poltrock, S. E., & Brown, P. (1984). Individual differences in visual imagery and spatial bility. Intelligence, 8, 93-138.

Raaheim, K. (1988). Intelligence and task novelty. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 4, pp. 73-97). Hillsdale, NJ: Erlbaum.

Raven, J. C. (1938-65). Progressive matrices. New York: The Psychological Corporation.

Robert, R. D. (1995). Speed of processing within the structure of human cognitive abilities. Unpublished doctoral dissertation, University of Sydney, Australia.

Robinson, C. S., & Hayes, J. R. (1978). Making inferences about relevance in understanding problems. In R. revlin & R. E. Mayer (Eds.), Human reasoning (pp. 195-206). Washington, DC: V. H. Winston.

Rosch, E. R., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7, 573-605.

Rumelhart, D. E. (1980). Understanding understanding. (Tech. Rep. 8101). San Diego: University of California, Center for Human Information Processing.

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. Cognitive Psychology, 5, 1-28.

Ryle, G. (1949). The concept of mind. London: Hutchinson's University Library.

Salthouse, T. A., Babcock, R. L, Mitchell, D. R. D., Palmon, R., & Skovronek, E. Sources of individual differences in spatial visualization ability. Intellligence, 14, 187-230.

Schank, R. C. (1978). Interestingness: Controlling inferences (Computer Science Research Report No. 145). New Haven, CT: Yale University.

Schank, R. C. (1980). How much intelligence is there in artificial intelligence? Intelligence, 4, 1-14.

Schank, R. C. (1984). The explanation game (Computer Science Research Report No. 307). New Haven, CT: Yale University.

Schunk, D. H., & Zimmerman, B. J. (Eds.). (1994). Self-regulation of learning and performance: Issues and educational applications. Hillsdale, NJ: Erlbaum.

Sharp, S. E. (1898-99). Individual psychology: A study in psychological method. American Journal of Psychology, 10, 329-391.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment (Research Report 97-9). Princeton, NJ: Educational Testing Service.

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

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. Science, 171, 701-703.

Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. S. Siegler (Ed.), Children's thinking: What develops? (pp. 325-48). Hillsdale, NJ: Erbaum.

Simon, H. A. (1979). Models of thought. New Haven, CT: Yale University Press.

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

Sincoff, J. B., & Sternberg, R. J. (1987). Two faces of verbal ability [editorial]. Intelligence, 11, 263-276.

Snow, R. E. (1978). Theory and method for research on aptitude processes. Intelligence, 2, 225-278.

Snow, R. E. (1980a). Aptitude and achievement. New Directions for Testing and Measurement, 5, 39-59.

Snow, R. E. (1980b). Aptitude processes. In R. E. Snow, P.-A. Federico, & W. E. Montague (Eds.), Aptitude, learning, and instruction: Vol. 1. Cognitive process analyses of aptitude (pp. 27-64). Hillsdale, NJ: Erlbaum.

Snow, R. E. (1989). Aptitude-treatment interaction as a framework for research on individual differences in learning. In R. J. Sternberg & R. Glaser (Eds.), Learning and individual differences: Advances in theory and research (pp. 13-59). New York: W. H. Freeman.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.), Mind in context: Interactionist perspectives on human intelligence (pp. 3-37). Cambridge, UK: Cambridge University Press.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), Educational measurement (3rd ed., pp. 263-331). New York: Macmillan.

Snow, R. E., Corno, L., & Jackson, D., III. (1996). Individual differences in affective and conative functions. In D. C. Berliner & R. C. Calfee (Eds.), Handbook of educational psychology (pp. 243-310). New York: Simon & Schuster Macmillan.

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. 47-104). Hillsdale, NJ: Erlbaum.

Spearman, C. E. (1927). The abilities of man. London: Macmillan.

Spilsbury, G. (1992). Complexity as a reflection of the dimensionality of a task. Intelligence, 16, 31-45.

Spiro, R.J., & Myers, A. (1984). Individual differences and underlying cognitive processes. In P. D. Pearson, R. Bar, M. L. Kamil, & P. Mosenthal (Eds.), Handbook of reading research (pp. 471-501). New York: Longman.

Stankov, L. (1988). Single tests, competing tasks and their relationship to broad factors of intelligence. Personality and Individual Differences, 9, 25-33.

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

Sternberg, R. J. (1980). Representation and process in linear syllogistic reasoning. Journal of Experimental Psychology: General, 109, 119-159.

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

Sternberg, R. J. (1986). Toward a unified theory of human reasoning. Intelligence, 10, 281-314.

Sternberg, R. J. (1990). Metaphors of mind: Conceptions of the nature of intelligence. Cambridge, UK: Cambridge University Press.

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

Sternberg, R. J., & McNamara, T. P. (1985). The representation and processing of information in real-time verbal comprehension. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 21-43). New York: Academic Press.

Sternberg, R. J., & Powell, J. S. (1983). Comprehending verbal comprehension. American Psychologist, 38, 878-893.

Sternberg, R. J., & Rifkin, B. (1979). The development of analogical reasoning processes. Journal of Experimental Child Psychology, 27, 195-232.

Sternberg, R. J., & Weil, E. M. (1980). An aptitude-strategy interaction in linear syllogistic reasoning. Jounral of Educational Psychology, 72, 226-234.

Sternberg, R. J., Powell, J. S., & Kaye, D. B. (1983). Teaching vocabulary-building skills: A contextual approach. In A. C. Wilkinson (Ed.), Classroom computers and cognitive science (pp. 121-143). New York: Academic Press.

Swiney, J. F. (1985). A study of executive processes in intelligence. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Tatsuoka, K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), Cognitively diagnostic assessment (pp. 327-360). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. (1997). Rule-space methodology. Unpublished paper. Princeton, NJ: Educational Testing Service.

Thorndike, R. L., & Hagen, E. P. (1993). The Cognitive Abilities Test. Itasca, IL: Riverside Publishing.

Thurstone, L. L. (1924). The nature of intelligence. New York: Harcourt, Brace.

Thurstone, L. L. (1938). Primary mental abilities. Psychometric Monographs, 1.

van Daalen-Kapteijns, M. M., & Elshout-Mohr, M. (1981). The acquisition of word meanings as a cognitive learning process. Journal of Verbal Learning and Verbal Behavior, 20, 386-399.

Vandenberg, S. G., & Kruse, A. R. (1978). Mental rotations: Group tests of three-dimensional spatial visualization. Perceptual and Motor Skills, 47, 599-604.

Wechsler, D. (1991). Manual for the Wechsler Intelligence Scale for Children--III. San Antonio, TX: The Psychological Corp.

Werner, H., & Kaplan, E. (1952). The acquistion of word meanings: A developmental study. Monographs of the Society for Research in Child Development (No. 51).

West, T. G. (1991). In the mind's eye. Buffalo, NY: Prometheus Books.

Whitely, S. E. (1976). Solving verbal analogies: Some cognitve components of intelligence test items. Journal of Educational Psychology, 68, 234-242.

Whitely, S. E., & Barnes, G. M. (1979). The implication of processing event sequences for theories of analogical reasoning. Memory & Cognition, 7, 323-331.

Winograd, T. (1972). Understanding natural language. Cognitive Psychology, 3, 1-191.

Winograd, T. (1975). Frames and the declarative-procedural controlvery. In D. G. Dobrow & A. Collins (Eds.), Representation and understanding: Studies in cognitive science (pp. 185-210). New York: Academic Press.

Wright, D., & Dennis, I. (in press). Exploiting the speed-accuracy trade-off. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), Learning and individual differences: Process, trait, and content determinants. Washington, DC: American Psychological Association.

Zimmerman, W. S. (1954). The influence of item complexity upon the factor composition of a spatial visualization test. Educational and Psychological Measurement, 14, 106-119.

Table 1

A Taxonomy of Semantic Relations

Semantic Type	Semantic Class	Example
Intensional	Class Inclusion	robin:bird
	Similarity	breeze:gale
	Attribute	beggar:poor
	Contrast	stutter:speech
	Nonattribute	harmony:discordant
Pragmatic	Event	tailor:suit
	Cause-purpose	virus:illness
	Space-time	judge:courthouse
	Part-whole	engine:car
	Representation	building:blueprint

Note. From Cognitive and psychometric analysis of analogical problem solving (p. ___), by I. I. Bejar, R. Chaffin, & S. Embretson, 1991, New York: Springer-Verlag.

Table 2

Attributes Hypothesized to Influence Performance on SAT Verbal Analogy Items

Processing Difficult Vocabulary	
01a	The ability to process low-frequency vocabulary in the stem.
01b	The ability to process low-frequency vocabulary in the key.
01e	The ability to process low-frequency vocabulary in the four distractors.
14	The ability to process longer words in the key.
15	The ability to process longer words in the distractors.
18	The ability to process longer words in the stem.
22a	The ability to process medium-length words in the whole item.
22b	The ability to process longer words in the whole item.
02	The ability to process vocabulary out of its usual context.
Processing Complex Concepts and Relationships	
03a	The ability to process multiple meanings due to semantic ambiguity.
03b	The ability to process multiple meanings due to syntactic ambiguity.
04b	The ability to process abstract concepts.
05	The ability to recognize relationships across semantic domains.
07	The ability to process a negative rationale.
08	The ability to process a complex rationale.
25	The ability to analyze and contrast concepts.
Deploying Background Knowledge	
09	The ability to process scientific topics.
10	The ability to discount the influence of emotive language.

(continues)

Note. The following interaction attributes were also retained in the final model: 14 ∞ 18, 01b ∞ 10, 14 ∞ 05, 18 ∞ 01b, 01b ∞ 22b, 01b ∞ 02, 15 ∞ 05, and 03b ∞ 22b. From "The Process of Identifying Cognitive and Linguistic Attributes Underlying Performance in the Verbal Domain: An Example with the SAT-V," by G. Buck, T. VanEssen, K. Tatsuoka, & I. Kostin, 1998, unpublished paper, Princeton, NJ: Educational Testing Service.