3. Toward a Cognitive Theory for the Measu rement of Achievement

Robert Glaser
University of Pittsburgh, glaser@pitt.edu

Alan Lesgold
University of Pittsburgh, al@pitt.edu

Susanne Lajoie
University of Pittsburgh

3 Toward a Cognitive Theory for the Measurement of Achievement

Robert Glaser
Alan Lesgold
Susanne Lajoie
Learning Research and Development Center, University of Pittsburgh

INTRODUCTION

Given the demands for higher levels of learning in our schools and the press for education in the skilled trades, the professions, and the sciences, we must develop more powerful and specific methods for assessing achievement. We need forms of assessment that educators can use to improve educational practice and to diagnose individual progress by monitoring the outcomes of learning and training. Compared to the well-developed technology for aptitude measurement and selection testing, however, the measurement of achievement and diagnosis of learning problems is underdeveloped. This is because the correlational models that support prediction are insufficient for the task of prescribing remediations or other instructional interventions. Tests can predict failure without a theory of what causes success, but intervening to prevent failure and enhance competence requires deeper understanding.

The study of the nature of learning is therefore integral to the assessment of achievement. We must use what we know about the cognitive properties of acquired proficiency and about the structures and processes that develop as a student becomes competent in a domain. We know that learning is not simply a matter of the accretion of subject-matter concepts and procedures; it consists rather of organizing and restructuring of this information to enable skillful procedures and processes of problem representation and solution. Somehow, tests must be sensitive to how well this structuring has proceeded in the student being tested.

The usual forms of achievement tests are not effective diagnostic aids. In order for tests to become usefully prescriptive, they must identify performance components that facilitate or interfere with current proficiency and the attainment of eventual higher levels of achievement. Curriculum analysis of the content and skill to be learned in a subject matter does not automatically provide information about how students attain competence about the difficulties they meet in attaining it. An array of subject-matter subtests differing in difficulty is not enough for useful diagnosis. Rather, qualitative indicators of specific properties of performance that influence learning and characterize levels of competence need to be identified.

In order to ascertain the critical differences between successful and unsuccessful student performance, we need to appraise the knowledge structures and cognitive processes that reveal degrees of competence in a field of study. We need a fuller understanding of what to test and how test items relate to target knowledge. In contrast, most of current testing technology is *post hoc* and has focused on what to do after test items are constructed. Analysis of item difficulty, development of discrimination indices, scaling and norming procedures, and analysis of test dimensions and factorial composition take place after the item is written. A theory of acquisition and performance is needed before and during item design.

Recent work in cognitive psychology is a good start toward a theory to underpin such measurement. Modern learning theory is taking on the characteristics of a developmental psychology of performance changes—the study of changes that occur as knowledge and complex cognitive strategies are acquired, and the study of conditions that can influence these transitions in competence. Achievement measurement must be designed to assess these performance changes. It must be cast in terms of development, or levels of acquisition, and must be informed by knowledge of sources of difficulty and facilitators of the growth of competence.

In essence, the theme of this chapter is that the measurement of achievement should be based on our knowledge of learning and of the course of acquisition of competence in the subject matters that we teach. We begin by sketching some findings of cognitive psychological research that have implications for achievement test design. We then give additional research examples from various subject-matter fields. A third section describes several analytic methods from our work that we think can be extended into new testing formats. Throughout, we emphasize the necessary inseparability of instruction and assessment and we consider, in this connection, the design of intelligent computer tutors, which require both an instructional and a testing capability. We conclude with suggested ingredients for a set of cognitive principles for achievement measurement.

COGNITIVE RESEARCH RELEVANT TO THE MEASUREMENT OF ACHIEVEMENT

A psychology of learning that can inform testing must address two central problems.

- First, we must understand how subject-matter knowledge is structured and how it changes with learning. That is, we need to understand the knowledge structure indicators of achievement.
- Second, we must understand how a particular piece of knowledge, a single performance rule, or a part of a procedure, becomes more reliable, flexible, adaptive, and automatic with practice. That is, we need to understand the performance indicators of achievement.

Knowledge Structures

A substantial body of research has been carried out on the knowledge structures that characterize experts in a domain. Unlike past research, which tended to concentrate on the prerequisites of learning, this work attempts to determine the nature of competent performance by examining the underlying cognitive structures of the expert. It therefore has the potential to reveal how processes are transformed in the course of a person's progress from the novice to the expert state of performance. The research shows that, compared to novices' knowledge structures, experts' knowledge structures are both wider and deeper. That is, they contain more concepts, with more detail about each and with more interconnections among them. However, since it appears that little can be learned without at least a partial theory to lend it coherence (Murphy & Medin, 1985), we can assume that the understanding even of novices is held together by at least a primitive organizational structure, or *personal theory.* Since personal theories evolve as more is learned, bootstrapping further learning, the type of theory a person currently holds for a domain can serve as an index of and basis for his progress in acquiring the knowledge of that domain.

Carey (1985) has studied the evolution of theories that children hold at different points in their cognitive development, concentrating on such domains as basic biology. Her work suggests that related reasoning and problem solving are greatly influenced by experience with new information. In her research on animistic thinking in children, she has shown how children's knowledge influences their conceptualization of being "alive" and how such a concept becomes more differentiated with time through school learning and experience in the world. For instance, a 5-year-old's knowledge of biological properties is organized in terms of the child's knowledge of human activities, whereas a 10-year-old's knowledge is organized in terms of biological functions. Asked whether worms or plants breathe, the younger children respond based on their experience of how human beings breathe and say "no," since they see nothing like a moving chest in a worm or plant; older children, who have been exposed to school-taught notions of respiration, are more likely to answer that worms and plants do breathe. Such abstract pervasive changes in the child's reasoning and learning abilities are repeated as knowledge is gained in various domains.

The theories that we have for domains that are acquired partly on the basis of everyday experience are extremely stable. They are not easily rejected in the face of counterevidence, especially if that counterevidence comes from a textbook or lecture. This has been noted in a variety of studies showing that students do not relinquish their naive views about force and motion even after a physics course (Champagne, Klopfer, & Anderson, 1980; Larkin, 1983), and Carey has made similar observations. For example, after a long interview in which many internal organs were discussed, children of ages 4 through 6 were asked what part of their body was most important. In spite of all the new information about internal organs, they tended to name an external feature such as nose, toes, or hair, something related to their self-observations of their activity, consistent with their activity-based theories about life.

Personal theories seem to have the same sort of resilience and ability to withstand counterevidence that are seen in scientific theories that are socially shared, and the abandonment of one personal theory for another may well be revolutionary rather than evolutionary, just as seems to be the case with scientific theories (Kuhn, 1962). The robustness of personal theories implies that in order to facilitate learning, i.e., transitions in knowledge structures, it is necessary to confront a person's theories with specific challenges and contradictions. Understanding how counterevidence and knowledge confrontation assist in the transition between levels of competence, we should be able to design instruction that will help students build from their existing repertoires. The research on personal theory building that will be most useful to an improved technology of measurement aims at (1) understanding the stages through which personal theories pass well enough to be able to detect them, and (2) being able to prescribe forms of instructional intervention that are appropriate to those stages.

Automaticity, Proceduralization, and Practice

John Anderson (1983) has developed a theory of the development of skilled performance based on the work of Fitts (1964). It divides the course of learning into three parts: the declarative stage, the knowledge compilation stage, and the procedural stage. Initial performance in a novel situation involves the operation of general strategies that use declarative knowledge to guide performance. *Declarative knowledge* refers to verbal rules or facts regarding a task. Accessing these bits of information may be a slow process in this stage, and the task procedure is slow, laborious, and requires conscious attention. A child learning to tie a shoe or to do subtraction, possibly verbalizing aloud, losing track if he or she is interrupted, is probably in the declarative stage of acquiring a skill.

The conversion of slow declarative knowledge into faster compiled procedures occurs in the second stage of acquisition, *knowledge compilation.* Knowledge compilation is analogous to compilation of a computer program, the translation of that program from an understandable verbal form to commands, in the form of bit patterns, that can be directly executed by the computer hardware. Compiled knowledge, like a compiled program, runs faster but at the cost of greater difficulty in modification. Compiled procedures are relatively automatic. They can be represented as systems of condition-action pairs called *productions,* which state an action to be performed whenever its associated condition, which is a specific memory state, is attained. A production normally proceeds without conscious control except when one of the conditions for productions is a goal state that has to be set consciously. For example, anyone who, after years of tying shoes, has tried to give verbal directions to a child realizes that even though he now ties shoes very efficiently, he no longer remembers the instructions he was once given and doesn't quite know what to say to the child. Knowledge compilation consists of two processes, *proceduralization* and *composition. Proceduralization* can be compared to the primary activity of compilation in a computer, but it is driven by experience-established connections rather than by a parsing process alone. If one successfully uses specific declarative knowledge in a specific setting, then the conditions at the time of the successful action are combined with the memory state needed to produce the action and stored in memory as a production. *Composition* takes place when two productions execute successfully in immediate sequence and thus become combined into a single production. It is similar to local optimization in a computer compiler. Proceduralization, then, is an automation process, whereas composition is an abbreviation process.

In Anderson's third stage the newly acquired productions become *tuned.* That is, they become strengthened, so that they prevail over other conflicting productions whose conditions may also match the same memory states, and their conditions for execution are more completely specified, through generalization and discrimination processes reminiscent of those described by Hull (1943) and Spence (1956).

A theory of skill acquisition such as this one has implications for test developers, since it identifies stages of learning and practice that are informative for instructional purposes. Lesgold (1984a; Lesgold, Rubinson, Feltovich, Glaser, Klopfer & Wang, in press) provided an example of how such a theory describes one aspect of the acquisition of expertise in medical diagnosis. Consider a resident who makes a faulty diagnosis during patient rounds. The attending physician may ask a series of questions that essentially walk the resident through the correct diagnosis. In spite of demonstration that this resident has the correct declarative knowledge, he or she is unable to organize the information in a manner that would lead to the correct solution because proceduralization has not taken place. The assessment of what knowledge needs to become organized and how to facilitate the proceduralization of such knowledge can greatly enhance instruction.

Of course, learning involves more than the proceduralization of declarative knowledge. Recall that we asserted in the previous section that new knowledge is acquired through the filter of one's personal theory. There are strong constraints on the verbal knowledge the student constructs from the things he experiences and is told. Consequently, a cognitive psychology of learning that deals only with what happens after the knowledge is already developed (albeit in fragile, declarative form) is not sufficient. Nonetheless, just as we have stated goals relating to the initial construction of knowledge, we can state some goals that a cognitive theory of measurement ought to have for dealing with practice and the automation of knowledge.[1]

A major emphasis in assessment should be to understand how the successive stages of learning, declarative, compiled, and tuned, manifest themselves in measurable performances. Combined with knowledge of how to foster progress from one stage to the next, this understanding will enable us to diagnose, to measure performance and to prescribe instruction based on those measurements. If we can develop both the capability to measure the stage of learning and the ability to assess which level of personal theory a student holds in a domain, then we should be able to make even stronger diagnoses.

Instruction might then guide the development of both the necessary declarative knowledge and its subsequent proceduralization and tuning. An emphasis on the conditions that foster the development of procedures, both simple and composite, will be necessary. Presumably, when teaching beginners we must build from their initial knowledge structures. This might be accomplished by assessing and using relevant prior knowledge, or by providing obvious organizational schemes or temporary models as scaffolds for new information. These temporary theories could be incorporated systematically into instruction. Such structures, when they are used, tested, and perhaps falsified by novices in the course of learning and experience, should lead to organizations of knowledge that are the basis for the more complete theories of experts. As well as assisting in developing theories, instruction can also systematically provide learners with the practice necessary for knowledge compilation and can encourage tuning by providing multiple contexts affording a chance to learn where certain procedures are applicable. This instructional emphasis should encourage discrimination and generalization of productions, leading to more robust, flexible, and efficiently

---

[1]We concede, of course, that these are not necessarily totally separate enterprises. How well certain components of a personal theory are automated may play a role in how resistant it is to being overthrown.

organized schemata[2] that allow the individual to perform appropriately under a variety of conditions. Acquiring expertise is to be seen as the successive development of efficient, tuned knowledge structures that facilitate the development of higher levels of competence.

A somewhat different approach to understanding the development of skilled performance is found in the work of Schneider (1984), whose research on the role of practice in training high levels of skilled performance also has several implications for assessment and instruction. We see Schneider's research as consistent with Anderson's theory of acquisition. Schneider, however, concentrates on an account of the development of automaticity, corresponding in particular with the knowledge-compilation stage. Environments must be designed, he asserts, to provide for the development of simple procedures. Once such procedures are developed, the sequence of instruction can facilitate a generalization between congruent procedures, fostering the composition and compilation process. He also suggests that construction of hierarchical knowledge structures can be facilitated by providing practice opportunities in multiple contexts, controlling the sequencing of levels of difficulty, and providing sufficient challenge and opportunities for success.

The areas of his investigations that could influence more diagnostic forms of achievement measurement to aid instruction include work on (1) identifying and training subskills rather than concentrating exclusively on total task instruction; (2) assessing levels of skill acquisition in order to facilitate the proper sequencing of instruction; (3) assessing individual differences in ceiling performance on a task; and (4) assessing the motivational aspects of learning the material under consideration.

In designing practice that is sufficient to produce high skill levels, Schneider suggests an emphasis on practicing consistent components of the task before practicing the task as a whole, even before the student understands a consistent mapping of required actions onto conditions. In other words, it is not just the amount of practice but also the focus of practice that matters. Schneider's approach places great importance on another goal for cognitive measurement theory: to formulate rules for deciding when a component skill is practiced enough to be integrated with other components to form a higher-order skill.

These contributions to theories of learning dealing with the role of practice have provided guides to the shaping of a diagnostic theory for the measurement of learning. For example, Schneider has suggested that goals for the level of proficiency to be attained must be set individually for different students, that different people show different cost-benefit functions for the marginal utility of

[2]Schemata are modifiable knowledge structures in memory that represent abstractions of experiences, including generic concepts, procedures, and situations (Glaser, 1984). They are used to interpret new instances of related knowledge (Rumelhart, 1975, 1981).

additional practice at different points in the course of learning. That is, not only the rate of learning but also the asymptote may vary from student to student. This poses another goal for diagnostic measurement, the assessment of potential for benefiting from particular components of an instructional program.

The bulk of the work in developing such a theory remains to be done. What we have now are some indicators of what a cognitive theory of measurement must be like. It must articulate with theories of learning and concentrate on shaping how we teach rather than whom we teach.

The Zone of Proximal Development

The idea of measurement techniques to measure the potential payoff of different instructional approaches is reminiscent of Vygotsky's theory of the zone of proximal or potential development, which was developed in the course of work on learning disabilities in the Soviet Union (Brown & French, 1979; Vygotsky, 1978). In this work, a distinction has been made between a child's actual developmental level (the level of mental functioning revealed in solo performance on a standardized test), and the child's level of potential development (the level of development that the child can achieve when offered certain forms of assistance). Both measures are considered essential for diagnosis and instruction. Vygotsky called the difference between these two levels the "zone of potential development," or "proximal development."

This zone of potential development is conceived of as an indication of learning potential. Thus, individuals with the same score on a mental ability test may vary in terms of their cognitive potential.[3] The relationship between assessments of the zone of potential and instructional strategies merits further research. The question is whether we could prescribe differential instructional treatment based on such a measure. Perhaps students with a large zone would do best being moved quickly through curricula, even skipping some units, while students with a small zone might require a slower, more complete treatment. In this manner, instruction might be prescribed so that learning neither lags behind potential nor pushes students beyond their capabilities. Presumably, motivation would be improved, too, if students were less likely to be overtaxed or bored. Extensions of Vygotsky's work (cf. Bransford, Delclos, Vye, Burns, & Hasselbring, 1986, and the chapters in Lidz, in press) represent an important step toward a cognitive instructional science of measurement.

---

[3]The distinction between crystallized and fluid intelligence (cf. Cattell, 1963) also seems to get at this issue.

Self-Regulatory and Metacognitive Skills

Metacognition has been defined in a number of ways across numerous subject domains and diverse populations. In general, though, metacognitive skills are generalized skills for approaching problems and for monitoring one's performance such as knowing when or what one knows, predicting the correctness of outcome of one's performance, planning ahead, efficiently apportioning one's time, and checking and monitoring one's thinking and performance (see Brown, 1978; Belmont & Butterfield, 1977; Borkowski, Cavanaugh, & Reichart, 1978; and Brown, Bransford, Ferrara, & Campione, 1983; for more extensive reviews). These skills, which act as control processes for cognitive performance, develop with maturity, and seem to be less developed in children with learning disabilities or those who are retarded. Brown (1978) suggests that these "executive processes" are a significant aspect of intelligence, since they determine when and where particular knowledge is used. Metacognitive skills are presumed to facilitate transfer of training to new situations.

In a sense, metacognitive skills represent, in part, performances that would be needed to realize the potential represented by the student's zone of proximal development. If we assess the zone of proximal development and attempt to specify and encourage the development of metacognitive skills, we are taking the first step toward trying to teach people to have larger zones. Thus, the movement toward task-analytic and instructional work on metacognitive skill is at the core of our aspirations for a technology of achievement assessment grounded in a cognitive instructional science.

Expert Performance

Understanding expertise is difficult because skillful performers appear to observe a set of rules that they themselves have difficulty verbalizing. This follows from the distinction made by J. Anderson between declarative and proceduralized knowledge (see p. 45), since experts can be assumed to have highly practiced repertoires of mental operations for tasks within their fields of competence. Seminal efforts to understand the nature of expert performance involved the study of skill in chess (Chase & Simon, 1973; de Groot, 1965, 1966; Simon & Chase, 1973). A series of experiments showed that the master chess player has a large repertoire of specific patterns that can be accessed in memory and quickly recognized. Chess expertise, to a large extent, is driven by rapid recognition processes that tap acquired structures of knowledge rather than by deep analytical thinking processes. Chess masters recognize the exact board situation they encounter and the strategies it entails; they do not excel by thinking ahead dozens of moves, as commonly thought; indeed they think ahead fewer moves than advanced players who are not yet at the master level (Chase & Simon, 1973).

Chess masters seem to have the ability to construct a qualitatively different representation of board positions than novice players, in terms of the aspects that they can immediately recognize and respond to (Chase & Chi, 1981). A similar phenomenon has been observed in more traditional school learning domains, such as physics, where highly competent performers also excel in developing an appropriate initial representation of a problem posed to them (Larkin, McDermott, Simon, & Simon, 1980; Simon & Simon, 1978). This representational skill allows the knowledgeable physicist to solve routine problems rapidly and without much conscious deliberation. An expert's representation of a physics problem tends to be organized around central principles of physics, whereas the knowledge of the novice is organized around more peripheral information such as the physical entities or objects described in the problem (Chi, Feltovich & Glaser, 1981).

The knowledge of experts and the mental representations they construct also include information regarding the application of what they know. In contrast, the novice's knowledge structure may be more loosely organized, containing the most centrally relevant information regarding the problem as stated but lacking the knowledge of related principles and their conditions of application. For this reason, novices may have more difficulty making inferences from the given problem statement. Their difficulties may be attributed to inadequacies of their knowledge bases as opposed to limitations on their capacities for carrying out problem solving processes.

In general, the competent individual can be described as having knowledge that is organized in a way that facilitates fast-access pattern recognition or encoding, greatly reducing mental processing load. These acquired knowledge patterns enable individuals to form an appropriate representation of the problem situation. The adequacy of the initial problem representation seems to be an index of developing competence, since the quality, completeness, and coherence of internal representations determine the efficiency and accuracy of further thinking. It seems appropriate then to consider the development of tests that will assess the learner's initial problem representations and level of knowledge organization.

Mental Models

Another research area with implications for a cognitive instructional theory of measurement deals both empirically and theoretically with the mental models that people construct in the course of solving problems. There are different kinds of mental models that are involved, and the implications for a theory of measurement may differ from one to the next. The "runnable" device or *qualitative process* model is perhaps the most important form. This type of model is a qualitative internal representation of a physical device along with a set of mental procedures for "running" that device, for simulating how the device changes as it operates. The *appearance* model is a related type, in which the person's procedural knowledge includes the ability to envision the appearance of a complex structure under various transformations. We discuss each of these below. In each case, we are concerned both with what is known about human capability and also with formal work attempting to specify what kind of modeling capability is needed to carry out various intelligent acts.

Qualitative Process Models. Qualitative physics is the effort to develop formalisms for representing the knowledge one can have about how things work (cf. de Kleer & Brown, 1984, and the entire issue of *Artificial Intelligence* in which it appears, "Qualitative Reasoning," 1984). One approach that has been taken is to represent each device in a system as a set of qualitative constraints (de Kleer & Brown, 1984). A device such as a resistor has qualitative constraints on it that are similar to Kirchhoff's current and voltage laws and Ohm's law. For example, the direction of change of current at one end of the resistor must match the direction of change in current at the other end, and the direction of change in resistance will be in the same direction as the change in voltage drop across the resistor. When devices are assembled into a system, the overall operation of the system can be envisioned by propagating the qualitative constraints of its components through the system. In a sense, then, running a device model is like solving a system of simultaneous quantitative equations.

It is very difficult to carry out this propagation mentally in real time. Experts tend to have highly practiced mental procedures for modeling a variety of common subassemblies of such systems. This makes them much faster, and at the limit more likely to succeed, in their mental modeling efforts. Further, because their modeling capability for routine situations is more efficient, they are more able to deal with novel variations from the routine. It should be possible to build tests of mental modeling capability by looking at relative speed and accuracy in an empirical progression of tasks such as (a) being able to state some of the constraints verbally but not being able to work with them, (b) having access to the most common, or classic, models in worked-out form, and (c) being able to modify these models to fit them to novel situations.

We have just begun in cognitive psychology to assess people's mental models, but this has been done in a few cases. For example, we can gather some of this information by asking people to predict the next state of objects in simple physics mechanics paradigms (McCloskey, Caramazza, & Green, 1980) or by asking them to describe simple electrical circuits (Gentner & Gentner, 1983, Riley, 1985). At LRDC, Jeffrey Bonar and his students have begun to develop environments in which subjects can make qualitative predictions of the effects of changes in a resistor network on various measurements in the network. We hope soon to be able to use such a capability to study possibilities for reliable mental model assessment.

Appearance Models. Another form of mental model is the *appearance model,* which represents how something looks or how it might look from various viewpoints. In studies we carried out on radiologists at differing levels of training, we realized that subjects varied in their ability to envision a patient's anatomy while looking at an x-ray picture. To assess their modeling ability, we asked them to draw, on the x-ray pictures, the contours of specific body structures (Lesgold, 1984a, Lesgold et al., in press). It was then possible to quantify performance by comparing the areas marked by the subjects with standard templates generated from expert protocols and other medical data. Measures such as proportion of template area covered by the subject's trace and proportion of the subject's trace that fell within the standard template region were computed. In this study, these measures were correlated with overall level of training and could be interpreted quite readily. Further, the subject's response (i.e., the tracing) could be input directly to a computer via various two-dimensional input devices, and the scoring done automatically.

Research on Acquisition of Subject Matters

It is in traditional school subject matters, of course, that achievement testing has had its widest application and most detailed development. Yet, perhaps it is school subject matters that most obviously demand a testing methodology that goes beyond normative scaling to become more relevant information for tailoring a student's instruction. As has been the case in testing methodology so far, different subject matters are likely to require different test item forms and perhaps even different overall testing approaches. We consider some of these in the sections that follow.

Reading. Progress has been made in understanding the nature of competence in reading, and there is beginning to be theory that might guide reading achievement test design. We can distinguish four reading processes that measurement should attend to. These four processes are: (a) word recognition; (b) accessing semantic word information; (c) sentence processing, and; (d) discourse analysis (see Curtis & Glaser, 1983).

A particularly important question is how the execution of one set of processes affects the efficiency of other reading processes. One component of the reading process that requires attention can affect reading comprehension by decreasing the amount of information maintained in memory and the amount of attention allocated to other processes. If, during reading, part of the thinking capacity is given over to word recognition, less capacity may remain for joining concepts that need to be interrelated in the reader's mind (Lesgold & Perfetti, 1978; Perfetti & Lesgold, 1977, 1979). That is, when word recognition is slow, comprehension processes become resource-limited (Norman & Bobrow, 1975), whereas faster recognition allows more effort to be directed to understanding what is read. In fact, poorer readers are generally slower at word recognition (Curtis, 1980; Lesgold & Curtis, 1981; Perfetti & Hogaboam, 1975).

Longitudinal research in the classroom, although difficult to implement, is a strong method for investigating the role of particular components (such as word recognition) over the course of learning a skill like reading. Rather than defining the development of learning in terms of grade level or age, the order in which subskills of reading are acquired can be specified directly if the same children are tested at different points in the course of their learning to read. This approach was used to observe the development of word recognition efficiency and its relation to comprehension skill development (Lesgold & Curtis, 1981; Lesgold, Resnick, & Hammond, 1985). Students were observed over a 4-year-period. Lesgold et al. (1985) examined student's reading efficiency in two reading curricula, one with an emphasis on word recognition training (phonics), and the other following a popular basal reading instruction program. Although no clear advantage was found for either curriculum, word processing speed measures did predict later reading comprehension in both groups. These results suggest that there are multiple approaches to developing reading comprehension but that automated word recognition is an important requirement for progress.

An interesting complication is that even though word recognition speed is the best predictor of reading achievement in the primary grades, as noted in the Lesgold study, listening comprehension becomes a better predictor thereafter (Curtis, 1980). This suggests that we do not yet have theories of the reading acquisition process adequate to support diagnostic testing. An adequate theory would have to account for the apparent fact that while word recognition ought to be the primary goal at the beginning of the curriculum, if a student is not doing well after several years, the focus needs to shift to comprehension skills. It will not suffice to use a checklist mastery approach, in which we have a schedule of subskills to be acquired, check off which ones the student has mastered, and diagnose that he should do the first thing on the list that is not yet checked.

A deeper understanding of how individuals retrieve word information can be used to guide assessment and instruction. There is a strong interdependency between ability to access the knowledge associated with words and overall comprehension skills. Three aspects of semantic retrieval capability seem to influence higher level processes: accuracy, flexibility, and fluency. (Beck, Perfetti, & McKeown, 1982). Understanding is not an all or none phenomenon; being accurate on one vocabulary item that uses a word does not necessarily mean that an individual fully understands that word. Items that reflect an individual's deeper knowledge of an item in terms of flexibility of usage in different contexts may be a more meaningful form of measurement. Qualitative differences in the levels of word knowledge can be assessed by presenting items that require specific and precise semantic discriminations (Curtis & Glaser, 1983). For instance, instead of a single word meaning question, a sequence of questions that reflect more detailed levels of understanding might be used, such as (a) *Which of* *the following synonyms best defines the word in question?* and (b) *Which of the following sentences uses the item correctly?* Contingent diagnostic testing sequences can be developed for individuals who vary in skill level. For example, if a student gets the simplest word meaning item incorrect, a subsequent question can be presented that gives the word in a context-providing sentence.

Thus, work to date suggests that diagnostic tests can be individualized to efficiently measure qualitative and quantitative differences in lexical/verbal knowledge. Efficient testing, in turn, might help make instruction more highly individualized. However, given the long history of difficulty in isolating multiple factors in tests of reading facility, it is clear that a sound theory of reading facility and its acquisition is needed before significant progress can be made. Recent efforts (e.g., Perfetti, 1985) seem a step toward such a theory.

Measuring comprehension skill raises a different set of issues. Understanding the sentences in a text requires prior knowledge. Knowledge of the topic or situation to which a passage refers, and knowledge of schemata, which are abstracted representations for situations and for discourse forms, can facilitate the understanding of passage content and its integration into existing memory organizations. Relevant schemata provide an interpretive framework for organizing the information mentioned in a text and for reading between the lines (inferring propositions which the author assumed did not have to be overtly stated, R. Anderson, 1978).

Hoepfner (1978) suggested that 10–20% of the items on reading comprehension tests assess schema-based knowledge. However, so long as these items are not recognized as dealing with a specific issue, their presence, through the natural selection processes involved in test construction and validation, does not provide any specialized diagnostic capability. If knowledge of specific schemata and prior knowledge of certain domains is a prerequisite to text comprehension and is not always sufficient, then comprehension items should be developed specifically to test for such knowledge. Another class of potentially useful items would test for inferential ability. Such comprehension tests would go beyond fact recall and test the subject's inferences based on the content of the passage. Presumably, these would be developed for discourse forms and topic domains for which the subject had previously demonstrated competence.

Another factor to consider when assessing discourse analysis is whether the examinee is having difficulty with comprehension in general or with reading in particular. This distinction generally is made by testing both reading and listening comprehension. However, it is important to note that comprehension in general is not wholly separable from reading comprehension; some argument forms simply cannot be presented orally, since they require too much temporary memory and therefore rely on the text itself as an external temporary memory.

To summarize, reading involves word recognition, lexical knowledge, knowledge of the forms in which discourses present information, and background knowledge for the domain about which any given text is written. Disciplined sequential testing strategies appear to have the potential for helping to isolate a student's reading problems to one or more of these areas. However, the current state, in which test items are developed without regard to a verified componential theory of reading acquisition and proficiency, and in which we lack the knowledge that would tell us that a specific item was measuring an identifiable skill component, does not permit tests to be used for detailed diagnosis.

Arithmetic. Arithmetic, like reading, is a basic skill that involves considerable procedural facility. It differs from reading in being dependent only upon a fixed domain of schematic knowledge (reading skill depends on schematic knowledge of the text topic). Because the schemata needed for arithmetic performances are less numerous, more refined theoretical analysis has been possible. A major program of research began when Brown and Burton (1978) developed computer models of children's substraction performance. They decomposed subtraction into very small procedural steps. Then they constructed degraded models, each of which contained all but one, or a small number, of the components of the full model. Since each degraded model made different performance errors, it was possible to assess a student's knowledge by trying to match the pattern of his answers to a set of subtraction problems with the pattern produced by one of the degraded or "buggy" models.

However, representing arithmetic errors as "bugs," deficiencies of a needed program step, was not sufficient to account for students' performance (Brown & Van Lehn, 1980, 1982; Van Lehn, 1983a,b,c). It became apparent that, while the bug analyses could account for the performance of students with systematic errors on any one test, a given student's bug patterns did not remain constant from day to day. The theory, which evolved from cross-sectional comparisons, did not transfer well to providing longitudinal accounts. Working with Brown, Van Lehn worked out a more complex theory which he called "repair theory." Its essence is that "bugs" do underlie failure of arithmetic performance but that students realize that they have reached impasses in their performance and make attempts to repair their incomplete procedures. When their knowledge of the basic conceptual underpinnings of arithmetic is solid, these repairs produce correct performance, and they manifest no stable error pattern. When their conceptual knowledge is inadequate, they are forced to invent ways of accommodating what they do know. For example, a student who doesn't know how to do regrouping (borrowing), when faced with a problem like

$$\begin{array}{r} 100 \\ \underline{-33} \end{array}$$

may answer "133," reasoning that there has to be a number in each column of the answer. If he can't compute $0-3$, then he computes what he can, $3-0$, instead. However, he knows he is likely to be wrong, so he doesn't stick with the specific strategy of always subtracting smaller digit from larger but rather tries other approaches from time to time.

The important thing to learn from this body of work is that some diagnosis may require longitudinal data, that the current knowledge of a person cannot always be determined in sufficient detail to suggest a specific approach to remediation or further instruction from looking only at current performance. A secondary lesson is that formal modeling approaches and the comparison of student performance to that of alternative models can be very useful strategies in designing new approaches to diagnostic assessment.

Word Problems. Quite a bit of work has been done on the kinds of problem solving that students are asked to do in school, such as the solving of arithmetic word problems. The general approach taken has been to attempt to specify the generic knowledge structures, or schemata, that subsume the knowledge needed to understand different categories of problem situations. As indicated above, schemata can be thought of as personal theories that we can test and revise. Learning can be thought of as being largely schema revision.

Riley, Greeno, & Heller (1983) have demonstrated that a small number of schemata can account for virtually all the arithmetic word problems that students are given in elementary school. Specifically, there are the following problem types, each of which requires different knowledge, i.e., a different schema:

- *Change.* Mary has i marbles and John gives her j more, so she has k in all. Any two of the three values would be given in the problem, and the student would have to find the third.
- *Combine.* Mary has i marbles and John has j. How many do they have altogether?
- *Compare.* Mary has i marbles and John has j. How many more does John have than Mary?

Further, Riley et al. suggested a developmental sequence for acquisition of these schemata. They found that problem schema type, rather than which arithmetic operations were required to solve a problem, was the best predictor of how early solution capability is acquired.

If solving a word problem requires knowing more than the arithmetic operations required to solve it, then the ability to diagnose student learning problems in arithmetic requires the ability to measure schematic knowledge, or to estimate it from the pattern of word problem types that a student can solve. Rather than simply looking at the total number of word problems a child solves, assessment procedures could examine or infer how students are representing the problem information. This form of measurement would indicate whether the student is having difficulty with the operations or with the semantic representation of the problem.

Writing. Recent work on the study of error in writing and composition has emphasized the identification of systematic misconceptions (see Bartholomae, 1980; Hayes & Flower, 1980; Hull, Ball, Fox, Levin, & McCutchen, 1985; Shaughnessy, 1977). While the same precision of error analysis that is seen in the mathematics work cannot be achieved in the writing domain, it is now clear that even students who write very poorly are following crude personal theories that they have formed for written communication. Systematic misconceptions or incomplete conceptions lead to errors, and can be detected from students' writing samples (Bartholomae, 1980). For example, poor writers may systematically mishandle verb endings, noun plurals, syntax, and sentence structure. If the current composition rules and schemata of a student can be determined, then presumably instruction can focus on specific efforts to move the student toward more appropriate understanding.

Like reading difficulty, poor written composition might, in principle, be due either to errors in general linguistic competence or to incomplete procedural rules for the specific medium, in this case writing. In order to rule out general linguistic competence as the problem in poor writers, Bartholomae had students read their writing samples out loud. In doing so he found that students, often unconsciously, corrected errors as they went along. This suggests that they have the general linguistic knowledge but do not have it, or cannot use it, in the specialized form needed to produce written products. As we refine our understanding of these procedural errors, we can better assess written composition and better develop individualized instruction aimed at repairing certain misconceptions and strengthening correct schemata.

Hull et al. (1985) used an extensive study of composition errors to develop computerized instruction in editing. Their software uses pattern matching techniques to assess systematic errors in writing and then helps students correct their own errors. Although this approach is still limited due to the complexity of error pattern detection in natural language texts, it has proven useful. By identifying errors, feedback can be provided to the learner regarding both the presence of errors and how to correct them. Highlighting of error regions in text-editor displays is used to help students learn to recognize and repair grammatical errors. Furthermore, instruction can be sequenced so that students can move from one level of skill to another, finding and correcting certain categories of errors and refining their own mental models. The integration of error identification with instructional remediation seems more promising as a diagnostic approach than are current tests of composition skill, although it remains unclear whether the breakthroughs are in diagnosis and individualized instruction or from increased understanding of the levels of competence in writing skills and of how learners can be assisted to acquire new knowledge given their current knowledge structures.

Scientific Concepts. An area in which much of instruction involves inducing change in students' schemata is science. There is now ample evidence that in certain cases where our environment provides a biased view of underlying natural processes, students tend to develop naive misconceptions that are extremely resistant to change (McCloskey, Caramazza & Green, 1980). For example, our everyday world, because of the friction effects of air and surfaces on which objects move, provides many experiences in which objects change velocity without being obviously affected by new forces. Thus, it is easy to conclude that force is required to sustain velocity, that velocity is proportional to force. After all, to go a constant speed in a car, you have to maintain constant pressure on the gas pedal. When students holding such misconceptions are exposed to formal physics instruction, they learn to solve physics problems that involve knowing that forces are proportional to accelerations, not velocities, but they do not generalize this knowledge to everyday life; they do not easily abandon their prior misconceptions. It seems unlikely that simply applying algorithms learned by rote will produce the needed learning. Thus, science instruction, like writing and arithmetic instruction, can be seen as involving diagnosis of a student's current schemata followed by efforts to move those schemata toward more expert form.

A methodology developed by Siegler (1976, 1978) is another promising approach to diagnostic measurement that is relevant here. Siegler assessed the underlying rule structure of certain cognitive performances and the progressive development of performance complexity in children. His "rule assessment" approach is based on two assumptions. The first is that human reasoning is rule governed, with the rules progressing from less sophisticated to more sophisticated as a function of age and learning. The second is that a way to assess these rule progressions is to develop diagnostic sets of problems that yield distinct performance patterns as a function of the rules a child knows. Just as with the arithmetic and writing research, this approach can determine what rules an individual uses in performing the task as well as what rules are common to various groups of individuals and age groups.

The first step in Siegler's procedure is to analyze the concept being studied. Through task analysis, one develops a first approximation of the condition-action pairing rules or specific rule knowledge that reflect competent performance on a task. Then, one attempts to characterize each known developmental stage as the presence of some subset of the final-stage rules or of rules with imperfect conditions or actions. The final analysis must be verified against actual children's performance. An acceptable set of rule stages has the property that each stage consists of only a small change, such as the acquisition of a rule or the elaboration of the conditions of a rule.

Siegler developed his rule assessment approach analyzing the performance of children on Inhelder and Piaget's (1958) balance beam task. The rules he identified involve understanding how balance is affected by the amount of weight applied, the distance of the weight from the fulcrum, the coordination of weight and distance, and finally how to compute the torques when necessary in order to choose the side of the scale with the greater value. These rules reflect a developmental progression in understanding the concept of balance. At this point, the methodology has not only provided a theoretical account of the capability being studied; it has also provided the basis for an instructionally diagnostic test. That is, one could identify a student's current stage and then proceed to teach him the rule or rule elaboration that enables performance at the next known stage.

Rule assessment approaches assume that conceptual development can be thought of as an ordered sequence of learned, partial understandings. If individuals learn concepts to various degrees of understanding and they develop understanding in a reasonably predictable fashion, the assessment of knowledge can be linked to appropriate instructional decisions. Rule assessment procedures, such as procedural analysis of arithmetic, error analysis in writing, and performance rules in scientific understanding, lead to diagnostic procedures that can provide deeper understanding of a subject matter that an individual brings to test performance. The concept of diagnosing test performance regularities at different levels of learning suggests a point of contact and possible integration of test theory, teaching practice, and the psychology of human cognition.

Technical Skill Development. Cognitive research on the assessment of technical skills is just beginning. During the past 2 years, the Learning Research and Development Center has been conducting a study of the feasibility of cognitive task analysis procedures for use in determining who should be placed in particular Air Force job specialties, how they should be trained, and how their performance should be measured. Our results have important implications for assessment and instruction. In addition to the traditional procedure of using aptitude tests for selection and using achievement tests at most for correlational evaluation of selection and adaptive instruction, we expect to assess achievement throughout the learning process. An important characteristic of our work is to compare the trainees who are most competent on the job with those least competent. When done at several stages in the progression from beginning apprentice to master, this provides a developmental view of the characteristics associated with success at different stages in the course of training.

To develop a cognitive task analysis of an area as broadly defined as an Air Force specialty, we took a job component sampling approach. We generated a representative sample of the tasks involved and examined their perceived trouble spots extensively. We were able to compare better to worse performers and to develop preliminary hypotheses about the different stages of performance in the course of the airmen's on-the-job experience. Our goals were to identify the procedural and conceptual knowledge required for job proficiency in using specialized test stations to isolate parts failures in aircraft navigation equipment. In these components, what flows through wires can be thought of as a simple signal with a small number of defining parameters, such as voltage.

We paid particular attention to how high and low performers differed in both conceptual and procedural knowledge. We also identified skills that should be automated in order for airmen to concentrate on higher order troubleshooting issues, and we developed tests of their automaticity. In addition, we tested for depth and organization of fundamental concepts and for understanding, in terms of functional systems on the aircraft, of the units that they were required to test. Through extensive protocol collection we observed the airmen's initial problem representation and the constraints used to arrive at solutions. Each of these assessment devices was guided by cognitive theory. Much of the remainder of this chapter, especially the following section, is shaped by our experiences in this project.

COGNITIVE RESEARCH APPLIED TO TESTING METHODOLOGY

So far, we have tried to highlight cognitive research on learning and expertise that is potentially relevant to building a richer theory of educational measurement. Such a theory, though, must have methods as well as principles. In this next section, we describe several methods that seem promising.

The Assessment of Flexible Problem Solving Skill

Assessing relatively general problem solving skills is quite a different task from assessing specific, algorithmic performance capabilities that are part of the domain being taught. We have only begun to work on this problem, but a few possibilities already present themselves, particularly with respect to the more strategic, or metacognitive, skills of problem solving. To give a sense of our work, we trace the history of our efforts to analyze the performance of electronics technicians when they attempt to troubleshoot complex electronic circuitry. The complex cases are of particular interest because they are the ones where metacognitive skills are needed to organize processes which, in simple cases, might automatically lead to problem solution.

In our first attack on this problem, Drew Gitomer[4] developed a troubleshooting task that involved detection of complex faults in the test station used by our subjects. As a first formative approach, he simply videotaped subjects attempting to solve such fault detection problems. He then examined the protocols (transcriptions of the tapes) and attempted to count a variety of activities that seemed relevant to metacognitive as well as more tactical aspects of problem solving in this domain. While the results, published in his thesis (Gitomer, 1984), were of great interest, we wanted to move toward a testing approach that was less

[4]At the time a graduate student at LRDC.

dependent upon skilled cognitive psychological training. That, after all, is one aspect of what test development is largely about—rendering explicit the procedures that insightful researchers first apply in their laboratories to study learning and thinking.

Our breakthrough came not so much from our psychological expertise but rather from our interactions with an electronics expert[5] who had extensive experience watching novice troubleshooting performances. He pointed out that it was not a big chore to specify all of the steps that an expert would take as well as all of the steps that any novice was at all likely to take in solving even very complex troubleshooting problems. That is, even when the task was to find the source of a failure in a test station that contained perhaps 40 cubic feet of printed circuit boards, cables, and connectors, various specific aspects of the job situation constrained the task sufficiently so that the effective problem space could be mapped out. This then created the possibility that we could specify in advance a set of probe questions that would get us the information we wanted about subjects' planning and other metacognitive activity in the troubleshooting task. For what is probably the most complex troubleshooting task we have ever seen, there are perhaps 55 to 60 different nodes in the problem space, and we have specific metacognitive probe questions for perhaps 45.[6] Figure 3.1 provides an example of a small piece of the problem space and the questions we have developed for it.

An examination of the questions in the Figure reveals that some are aimed at very specific knowledge (e.g., *How would you do this?*), while others help elaborate the subject's plan for troubleshooting (consider *Why would you do this?* or *What do you plan to do next?*). Combined with information about the order in which the subject worked in different parts of the problem space, this probe information permits reconstruction of the subject's plan for finding the fault in the circuit and even provides some information about the points along the way at which different aspects of the planning occurred. In fact, we went a step further and also asked a number of specific questions about how critical components work and what their purpose is.

After reviewing the protocol, we developed six scales on which we scored each airman. Each of these scales could be further subdivided into subscales to permit more detailed and task-specific issues to be addressed. The six scales were titled *plans, hypotheses, device and system understanding, errors, methods and skills,* and *systematicity*. Table 3.1 gives two examples for each scale of the items for which points could be earned (in the error scale, more points means more errors and thus is a lower score).

---

[5]We are grateful to Mr. Gary Eggan for his many insights in this work.

[6]Debra Logan and Richard Eastman have been refining this technology in our laboratories (Logan & Eastman, 1986), and we expect that a more detailed account will be published by them at a later date.

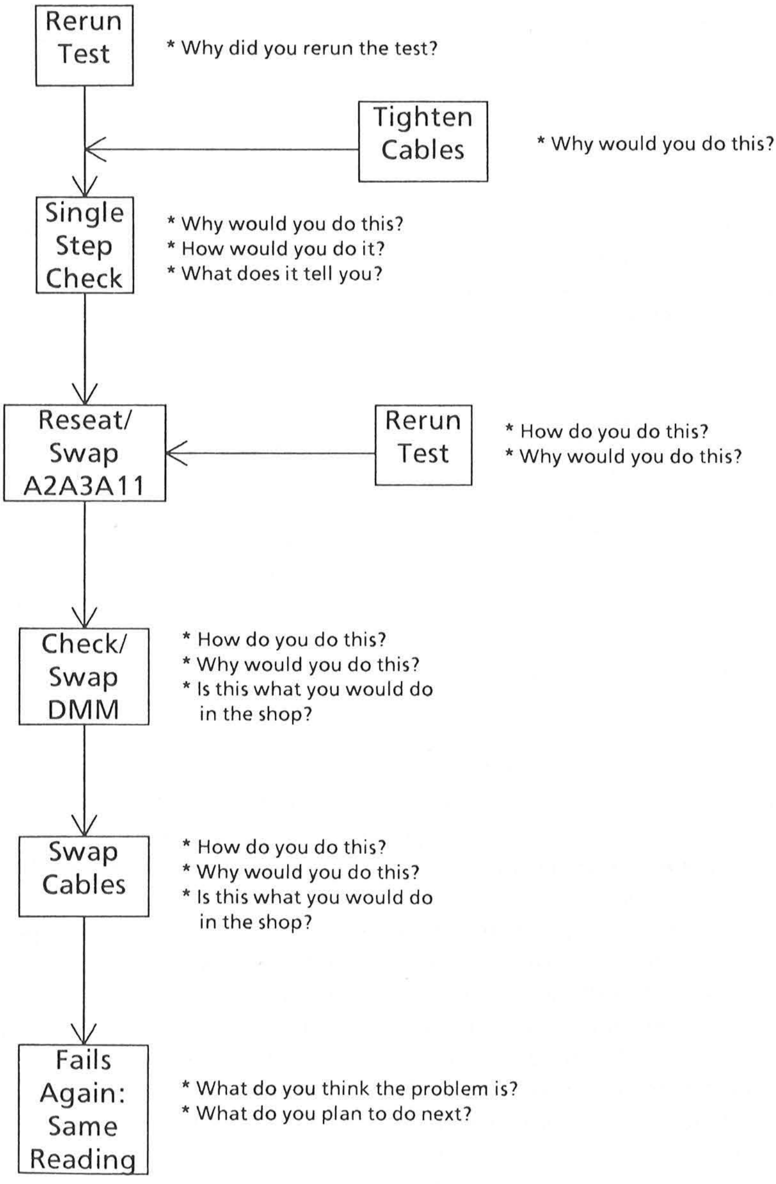

FIG. 3.1. Problem space map to guide probed protocol gathering.

TABLE 3.1
Examples of Criterion Questions Used to Score Protocols for Each of the Six Problem-Solving Scales

- PLANS
  - Extend and test a card.
  - Trace through the schematic of an individual card
- HYPOTHESES
  - There is a short caused by a broken wire or a bad connection.
  - The ground is missing from the relay.
- DEVICE AND SYSTEM UNDERSTANDING
  - Understanding and use of the external control panel.
  - Understanding of grounds and voltage levels in the test station.
- ERRORS
  - Misinterpreting/misreading the program code, called FAPA, for a test that the test station carries out under conputer control.
  - Getting pin numbers for a test wrong.
- METHODS AND SKILLS
  - Schematic understanding: Ability to interpret diagrams of relays, contacts, coils.
  - Ability to run confidence check programs.
- SYSTEMATICITY
  - The subject returns to a point where he knew what was going on when a dead end is encountered.
  - The path from the power source is checked.

Plans was a count of the number of plans mentioned by the subject during his problem-solving efforts. Any time that the subject entered a new part of the problem space, we prompted for a plan, but the lower skill subjects, especially, often did not have one. That is, they more or less randomly acted until a plan or hypothesis came to mind. A count was kept of the number of hypotheses offered by subjects at various points in their work. Again, subjects were prompted for hypotheses at the predetermined boundary points between regions of the problem space. The high-skill group entertained more hypotheses, which is what we would expect given that even they are at intermediate skill levels. True experts could be expected to have a more constrained set of probable hypotheses (cf. Benbassett & Bachar-Bassan, 1984; Lesgold et al., in press).

The *device and system understanding* scale was based on specific questions that were put to the subjects after they had performed the troubleshooting tasks. We asked a fixed set of questions about each of the components of the test station that played a role in the problems we had posed. These questions probed for knowledge about how the component worked, what role it played in the test station, what its general purpose in electronic system was, and what it looked like. The *errors* scale was simply a count of the number of incorrect steps taken by the airman in trying to troubleshoot the system. The *methods and skills* measure tallied which of the procedures needed to carry out the troubleshooting of the test station were successfully demonstrated by the subject. Finally, the *systematicity* measure consisted of a set of relatively broad criteria gauging the extent to which troubleshooting proceeded in a systematic manner rather than haphazardly or without a sense of goal structure.

With these scales, it is possible to provide a reasonable account of the components of performance. That is, there were no statements or behavior sequences of the airmen that could not be counted on one of our scales. This demonstrates that it is feasible to measure directly such complex cognitive performances as fault isolation in massive circuitry. By careful planning and the use of expert consultants with on-the-job supervisory and training experience, it is possible to develop measurement approaches that can help pinpoint a technician's stage of acquisition and, consequently, the level of further training needed. The approach is still rather expensive, but we feel that it is rapidly reaching the level of rigor associated with good experimental technique. Given its potential for more direct ties to theory by sharpening the criteria for the various scales used, it compares quite favorably with traditional approaches, which involve multiple-choice questions about somewhat simpler and less job-linked knowledge.

Gaining Objectivity and Simplicity

While cognitive psychology provides much guidance on what tests should be measuring, it has not so far contributed much to the technology of low-cost measurement. This optimization of cognitive measurement methods is critical to bringing cognitive science to bear on testing. If cognitive measurements cost two or three orders of magnitude more, they will not be used, even if they are the best alternative. We need to start searching for a middle ground between the overly-constraining 5-foil multiple-choice item and expensive verbal protocol procedures such as that just described. The multiple-choice formats currently used present two problems for us. On the one hand, they do not allow all of the responses subjects are likely to make to be included as alternatives, so great care is needed to understand how the range of possible student approaches will map onto a restricted set of possible answers. On the other hand, they tend to "give away" some aspects of the solutions to problems. That is, they can only be used where recognizing that one has a correct solution is sufficient performance. Below, we discuss some new ways to extend simple forced choice methods into the realm of complex cognitive activity. These approaches come closer to being "direct readouts" of knowledge and thus are more useful in building a representation of a person's cognitive capabilities.

Hierarchical Menus Methodology[7]

Computer-based menu systems offer the opportunity for extending the multiple-choice technology almost infinitely. Traditional multiple-choice tests require selection from a small set of alternatives. More elaborate alternatives have not generally worked well, probably because they impose a greater verbal processing load on the subject, who must keep in mind too much information at once in order to use them well. What the computer offers is the possibility of complex, choice-specific follow-up to individual items without placing any new test-taking skills demands on the subject.

In a sense, all computerized adaptive testing involves contingent sequencing of multiple-choice items. However, in existing adaptive procedures, the sequencing is not based on the content of the items, but rather on their classification into pools of different difficulty levels and different subscales. The same basic idea can be used to develop a cognitively oriented adaptive questioning procedure that is driven by propositional inferences rather than by statistical inference. The approach can best be understood through an example.

Suppose we wanted to know whether a student knew how to compute the mean of a set of numbers. If we simply want to determine whether he has this skill completely or not, we could make up simple multiple-choice items, such as the following:

> The mean of the numbers 1, 3, 4, 10, and 15 is (a) 6.6; (b) 33; (c) 5; (d) 4; or (e) 15.

If the student chooses *a*, then he is correct. However, we can learn from the errors, since *b* is the answer one would get if every step but the final division were carried out, *c* is the count of the numbers, *d* is the median value, and *e* is the maximum. However, we cannot actually see how the student tried to represent and solve this problem, so we don't have any ability to separate correct knowledge that is not sufficiently practiced from incorrect knowledge. It would be useful to be able to give a test that objectively and replicably recovered the actual content of the student's performance on this problem. From that, we could construct remediation, additional practice activity, or additional new instruction that might serve the student better.

Jeffrey Bonar in our laboratories has developed an approach to computer-based programming instruction (called BRIDGE) that can do this. The approach is based on a hierarchical menu scheme. The subject is asked a broad question

[7]This section is very much inspired by an approach Jeffrey Bonar has taken to the development of menu alternatives to natural-language input for computer-based instructional systems.

that can be answered by choosing one of several alternatives. The choice of alternative also determines the nature of followup questioning. Finally, the whole process can be repeated several times to allow specification of a multistep solution to a problem. The methodology rests upon a combination of a full analysis of the task to be performed by the subject and a set of protocols of people trying to do the task.

In fact, one of the tasks that Bonar has worked with is writing a computer program to compute the mean of a set of numbers. In most current computer languages, this is done more or less as follows:

```
Set a counter to zero
Set a sum register to zero
Read the first number to be included in the average
While the current number is not the termination code do
    Increase the counter by 1
    Add the current number into the sum register
    Read the next number
If the counter is not equal to zero
    Divide the sum register value by the counter value and
    Place the quotient in the sum register
Print the sum register value
```

Unfortunately, no questioning scheme based on the correct algorithm will work. This is because there is a very different way that students think about this problem before learning computer programming. A student asked to describe what he would do will say something like this:

> Get the first number and write it down unless it is the stop code. Then do this over again for each new number. When you reach the stop code, count the numbers and add them up. Divide the sum by the count, and that's your answer.

What Bonar has done is to give problems like this to a large sample and then analyze the procedures they wrote down. After completing these analyses, he was able to create a hierarchical menu system that allows students to specify their algorithms in ways that do not make it appear that they understand more than they really do. For example, it distinguishes between formally specifying an iterative process and simply stating that some steps must be repeated without being explicit about which steps or about the condition for ending the iterative loop. Recently, the same problems were given to a set of enlisted military personnel. Their performance could be fully accommodated by the set of options originally generated in response to protocols from college students, so the method seems robust.

Figure 3.2 shows the computer screen at a point where this interrogation is underway. Bonar's hierarchical menus allow specification even of very complex algorithms because of the followup questioning capability, and because each step already specified is displayed on the computer screen. For example, if the subject picks the REPEAT option from the menu, he is then prompted to indicate which items should be repeated. He is also asked to specify how to decide when to stop the repetition. The options for stopping the repetition can include simple tests, tests based on the result of procedures, and implicit tests (do it for every member of a defined set). We see this approach as an important step toward the building of cognitively oriented diagnostic tests. To some extent, Bonar's work bears out our beliefs. He is using this type of menu capability in a programming tutor as the basis for coaching the student toward the specification of precise algorithms and finally the specification of an actual computer program.

To summarize the last sections, we see great promise in such computer-based approaches to testing. These approaches will go beyond and build upon many of the best intuitions of current test item writers and computerized adaptive testing researchers. One new aspect of the work will be deeper, more interactive interrogation via menu-based systems supported by graphics and other verbal-load-reducing aids. A second will be processing of the test responses that is driven by logical inference from knowledge of the domain and knowledge of how students learn in the domain rather than only by statistical inference based on normative item difficulty and internal consistency data.

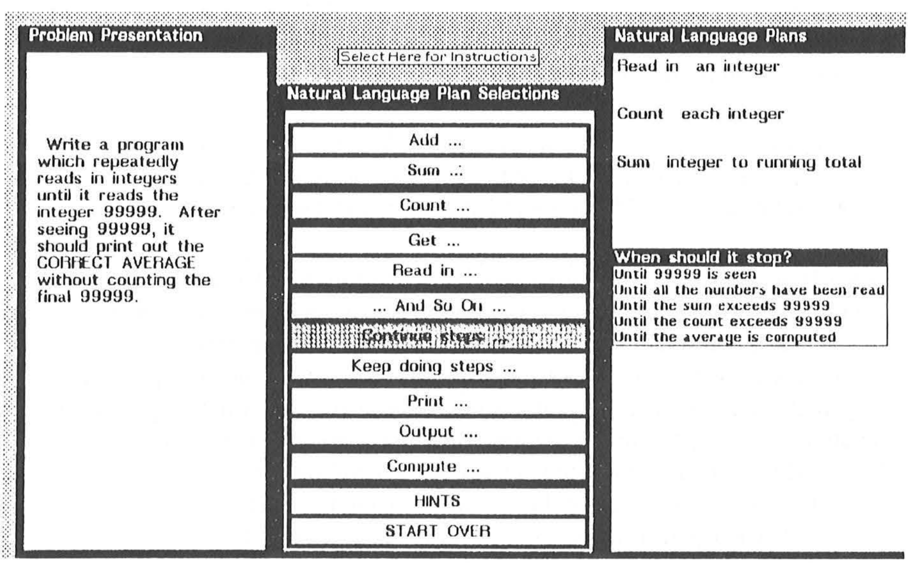

FIG. 3.2. Menu illustration from Bonar's BRIDGE programming tutor.

The Procedural Ordering Task:[8] Measuring Sensitivity to Constraints

In a number of areas of skilled performance, the range of acceptable responses to a situation is often quite wide and the constraints often are rather abstract. This makes the measurement of competence quite difficult. Consider a case recently encountered by Richard Eastman, one of the research assistants in our group. We were attempting to measure how well airmen could carry out the task of reassembling a complex part of a jet engine after overhaul. The procedure we thought likely to work well was an ordering task. We would take the actual steps of the reassembly task directly out of the manufacturer's manual, print each step on a card that also included a picture of that step, and see if the airmen could sort the cards. This seemed very straightforward until we found that there was no correlation between performance in this task and our other indices of expertise.

This led Eastman to interview several known experts. When he asked them why they had not used the ordering indicated in the documentation, they pointed out the underlying constraints on successful performance and also clarified how the procedures specifically listed in the manufacturer's documentation were inefficient ways to satisfy those constraints. The constraints all involved preserving the calibration of an information pathway between the engine and a gauge in the cockpit of the plane. The manufacturers, wanting to get foolproof instructions written quickly, had never included the constraints in the documentation. Rather, they had chosen to write a set of instructions that, while inefficient, would keep them out of trouble, since they happened to preserve the constraints.

Making up test items based only on the documentation would have failed to capture the full range of knowledge that is included in this particular brand of expertise.[9] Further, it was necessary to have deeper understanding ourselves than could be gleaned from the printed materials that would have driven most standard test-writing exercises in this area. Finally, we should note that a particular expert's favorite alternative to the textbook method could still have two forms: rote knowledge and conceptually-deep knowledge. That is, a mechanic might simply be following someone else's approach rather than the book's but still understand neither.

To get around this problem, Eastman cleverly designed a family of sorting tasks that varied in which devices were already fitted to the engine and which still had to be attached. Knowing the constraints, an expert would sometimes be able to leave a piece of the system attached rather than having to remove everything and start from scratch. However, to preserve calibration, it is also sometimes

---

[8]This task was developed by Richard Eastman as part of the cognitive task analysis project we conducted for the Air Force Human Resources Laboratory.

[9]We take no position on whether such instructions should be followed in every case simply to preserve a disciplined approach to maintenance—that is an Air Force issue.

necessary to remove an already-installed component which cannot be finally calibrated until some other device has been attached. Thus, unless the airman has learned dozens of rote variations of the assembly procedure, he will not be able to give optimal and still safe performances in these varying situations. Further, it is possible to model the sorting performance for such a family of tasks.

When Eastman looked at the pattern of errors for a small sample, he found evidence consistent with two hypotheses. First, the errors made by rank beginners tended to involve orderings of activity that are physically impossible (e.g., one can't get device A attached if device B is already attached and in the way). Later in the course of acquisition, the errors were more deeply conceptual (you could get all the parts together that way, but the information pathways would not be calibrated). We have experimented with computerized presentation of this kind of ordering task, and it seems quite straightforward.

In our preliminary efforts, we present the steps of a procedure as a menu on a workstation equipped with a pointer device called a mouse. When the subject points to an entry in the menu, the step is illustrated on a high-resolution screen, minimizing unnecessary verbal load. Pressing a button on the mouse causes the item to which the subject is pointing to be added to an ordered list of steps. A simple arrangement allows the subject to rearrange that ordering until he is satisfied with it. He then points to a box labeled "Done" and presses the mouse button. By having a list of constraints of each type, the computer can then report scores for physical adequacy as well as functional adequacy of the proposed orderings. The testing can be repeated with different starting scenarios to establish the character of the subject's knowledge.

The ability to produce or alter graphical displays is an important new capability. We see many possibilities for new test forms that involve pointing to locations or tracing regions in graphic displays. This approach, and others that provide more direct expressions of subjects' knowledge (as opposed to the ability to verbalize about knowledge), will be helpful to the development of a cognitively based testing technology for technical training.

INTELLIGENT TUTORING SYSTEMS: LABORATORIES FOR INTEGRATING TESTING AND INSTRUCTION

At the Learning Research and Development Center, we have embarked on a substantial program developing intelligent computer-assisted instructional systems—expert systems for teaching and training. We have done this for at least two reasons. First, we feel that sufficiently facile tutoring systems can help teachers improve their teaching skills as well as directly tutor students. Second, we see the expert instructional system as a primary laboratory for testing emerging principles for measurement and instruction. Since expert systems are driven by explicitly specified knowledge, they are direct empirical tests of the hypotheses embodied by that knowledge. In this section, we concentrate on possibilities for testing hypotheses that involve the measurement of performance.

Tutor Architectures and Adaptive Instruction

Intelligent tutoring systems are the ideal laboratory for investigating new assessment techniques, because the fundamental activity of such tutors is driven by assessment of individual student knowledge. Further, because such tutors embody explicit representations of theoretical assertions about learning, they are perhaps the least confounded forms of experimental treatment for empirical investigations of new ideas about assessment and instruction. An explicit set of roles must be programmed into any intelligent tutor. At times the tutor plays the role of diagnostician, trying to decide what the student does and does not know. At times, it plays the role of strategist, trying to decide how to respond to the student's weaknesses by tailoring instruction. At times it plays the role of colleague or foil, interacting with the student as coach or advisor, or even as game opponent.

In some intelligent tutors, such as WEST,[10] separate major segments of the program correspond to these separate roles. There is an expert modeler, a student modeler, a set of issue analyzers that determine differences between the student's performance and the ideal and blame those differences on particular student shortcomings (see Fig. 3.3). There is also a module that plays the game with the student, and a module that uses a prioritized set of pending issues (things the student should be taught) to decide how and what to coach.

While the roles to be filled by the WEST tutor are very explicit, the curriculum structure is much more implicit. There is no explicit statement embodied in the program that makes it clear what any given student will be taught by the tutor. Rather, a variety of considerations interact to determine what the student is taught. This can pose two types of problems. First, different schools may have different emphases. For example, one school might favor arithmetic instruction over refinement of gaming strategy, while another may emphasize the metacognitive skills involved in successful play. Second, we want to include knowledge about the course of learning that is not reflected in a model of expert performance alone, nor even in the sorts of tutoring principles currently found in programs such as WEST. The first problem is not very severe; one could change the

[10] WEST is an intelligent tutor developed by Richard Burton and John Seely Brown (1982). It is based on an instructional game developed by Bonnie Seiler. The game is a variation of Chutes and Ladders (a children's game) in which the student must develop an arithmetic expression instead of rolling the dice to generate a move. Three randomly generated numbers are presented to the student, who can then move his game piece as many squares as are represented by the value of any one arithmetic expression he can specify that uses only the three numbers. WEST provides advice, or coaching, to the player, on arithmetic issues, game strategy, and the manipulation of arithmetic expressions.

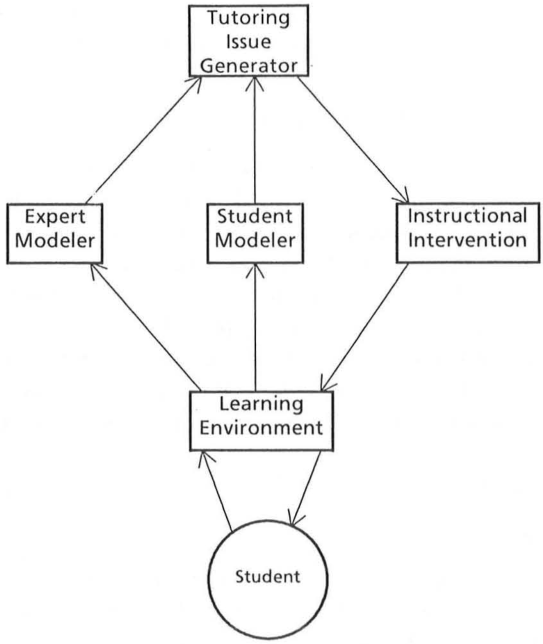

FIG. 3.3. Major components of a tutor such as WEST.

prioritizing rules used by WEST to choose which of several potential tutoring issues is given priority. The second issue, the need to reflect knowledge about the course of acquisition for a specific domain of expertise, is more serious and will be discussed below.

Representing Curriculum Knowledge As Well As Domain Knowledge

In order to separate knowledge about how and what to teach from the expert knowledge that represents the goals of instruction, a group of us[11] at LRDC have been experimenting with a new architecture for tutors. In it, the sections of the program correspond explicitly to the lessons in a curriculum, and the roles that the tutor must fill are defined in a distributed manner, as part of the content of each lesson's program. This curriculum knowledge "layer" is separate from an expert domain knowledge layer. This basic intelligent systems design approach has been proven in the speech understanding research of the past decade or more,

[11]The original ideas came from Jeffrey Bonar (1985), who continues to paly a central role with us in developing this new approach. Other important contributors have been Paul Resnick, William Weil, Cindy Cosic, and Mary Ann Quayle.

and we are adapting it to the task of building an intelligent tutor. It permits direct expression of hypotheses about instruction and diagnosis in the knowledge base from which an intelligent system develops specific lessons. Thus, we think it has promise for a technology and science of instruction.

The approach that we have been taking is to specify the knowledge relevant to tutoring as consisting of three types. First, there is domain knowledge, the content to be taught. Second, there is curricular knowledge, the division of the domain to be taught into a hierarchy of instructional subgoals. Finally, there is aptitude-related or metacognitive knowledge, the tailoring of the course of instruction to suit individual student needs. The three knowledge types are shown in Fig. 3.4, and we shall discuss each in turn.

Whatever it is that we want the student to know after he has been taught, we can represent it as a network of concepts connected by predicates. Certain groupings within such a network are organized into schemata, which contain the

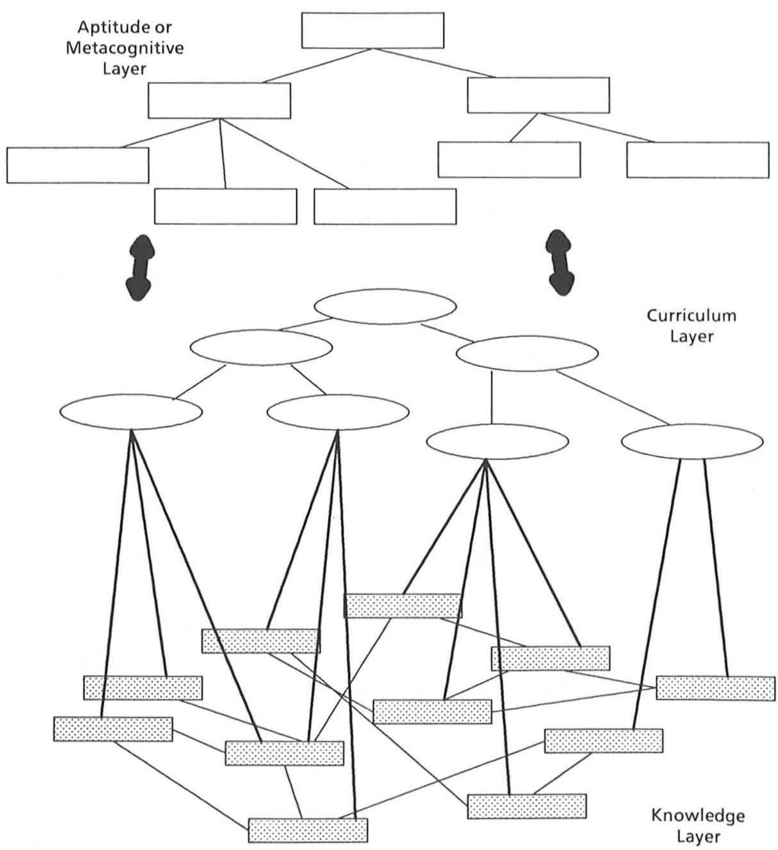

FIG. 3.4. Layers of a tutor's knowledge base.

knowledge, both procedural and declarative, needed to deal with particular broad generic problems. While this higher level organization is central to expert performance, the knowledge one must acquire to become an expert is not neatly separable into clusters that can be taught or measured independently. Expert knowledge has the character that each part one might want to teach or measure is somewhat depedent on other parts which may well not have been taught yet. Thus, any plan for building expertise must include attention not only to the skills an expert has but also to the sequencing and forms of instruction used to build those skills.

An important source of the needed structure is an understanding of how the expertise being taught is learned. That is, particular pieces of knowledge are propaedeutic and thus should be taught first. Other pieces of knowledge seem to be acquired more readily if they are taught only after some particular prerequisite has been taught first. While many people believe that prerequisite structuring is a property of the domain knowledge alone, a central assertion of the cognitive psychology of learning is that prerequisite structuring depends also upon what sorts of knowledge bundles lead to what sorts of learning and what sorts of capability for further learning. Two examples may help clarify what has just been stated.

The first of these examples has to do with acquisition of substantial understanding about the physical world, including scientific understanding. As noted above, Susan Carey (in press) has noted that the course of development for certain kinds of scientific knowledge recapitulates the sequences of theories that have developed in various sciences. We can imagine that some personal theories are better scaffolding for one kind of later learning than for another just as has been the case for scientific theories. For example, the alchemy theories were a very poor foundation for a quantitative chemistry that could be related to an emerging physics. On the other hand, they may have been better suited to the fostering of a materials science. Indeed, much relevant descriptive knowledge of the properties of different materials was temporarily lost when modern chemistry overthrew alchemy.

The moral in this discussion is that an understanding of the particular aspects of expertise that are most important, along with understanding of the course of acquisition for that knowledge, is the basis for splitting the knowledge domain up into bundles or lessons to be separately taught and separately tested. Those lessons, or curriculum subgoals, can be thought of as constituting a separate layer of process and control in an intelligent tutor with connections downward to the knowledge layer. The particular bundlings that are most optimal will be determined by analysis of the course of learning.

A second example may help clarify this viewpoint. Consider the learning of arithmetic. We want children to be able to carry out all four arithmetic operations on numbers of any size. How do we decide whether to teach addition first, proceeding only with addition until even the largest numbers can be added easily, or to teach two or more operations on small numbers and then recycle through the operations again with progressively larger quantities? Currently, schools do the latter, because it turns out that certain parts of each operation reinforce parts of other operations. However, some have argued that a modest movement toward providing large-number problems earlier, which might mean working with fewer operations initially, would clarify certain matters that now seem to cause trouble (for example, you can't understand how to borrow across zero, as in 403−224, unless you have had enough experience with 3 and 4-digit numbers to really establish your understanding of place value). Only when we know enough to specify the order of instruction can we be sure we understand how to handle the measurement problem with respect to diagnosis.

We propose that yet a third layer is needed in the knowledge base for a tutor. We call this the aptitude or metacognitive layer. Basically, this layer is concerned with individualizing the course of instruction to suit different students' capabilities. We presume that this layer sits on top of the curriculum layer, observing the student's progress and tuning the system's performance to optimize that progress. Thus, this layer is concerned with capability that is needed in order to become expert in a domain but is not actually part of the domain.

The optimization of student progress can occur in two ways. The system can either adapt to aptitude differences, or it can teach the skills that constitute aptitude. Adapting to aptitude might involve no more than changing the level of risk taking for the instruction. For example, if we have a student who learns with great facility, perhaps we should be less concerned if he misses a few problems in an early lesson, since he will probably pick up what he needs implicitly even if we give parts of the knowledge minimal explicit treatment. On the other hand, when we have a slow student, we may not want to take too many risks. Rather, we may not advance to a new lesson until the prerequisite lessons have been very well learned. A related form of aptitude optimization involves variation in the amount of support provided to a student in learning environments which foster discovery learning. For example, one might have a simulation program for a physics course that allows a variety of mechanics experiments to be simulated on the computer screen. For some students, very slight prompting might work very well, e.g., *See what you can figure out about the relationship between force and acceleration.* Other students with less-well-developed skills for exploratory learning might be led much more carefully through a specific set of experiments and prompted to specify what they had learned from particular experiences.

A different approach to aptitude is to diagnose the specific skill weaknesses that make some people learn better or faster than others. Under this approach, metacognitive skills are taken to be part of every domain of instruction, though they may be adapted specifically to each knowledge domain. The notion that we have in mind is one of observing the course of a student's progress in learning, inducing that certain aspects of metacognitive skill might be weak, and then establishing the teaching of those metacognitive skill components as high priority instructional subgoals. For example, the tutoring system might observe that the student never takes advantage of opportunities to conduct his own experiments in a simulation environment and respond by suggesting occasions when some experimentation would allow the student to test his understanding. Systems that begin to do this are being developed in our laboratories. With this approach, the metacognitive layer modifies the content of the curriculum layer by inserting additional curriculum subgoals. This can be of particular importance in instructional systems that attempt to foster discovery learning, since the student whose skills are insufficient for making discoveries is not well served otherwise.

Diagnosis With Curriculum Object Structuring

The tutor architectures we have been producing are based on an object-oriented programming approach. In such an approach, programs consist of independent modules (or objects). Each module contains a set of variables, the module's knowledge base, and a set of methods for responding to input messages. Control in such a system involves sending a message to one object which then carries out the method signaled by the message. Some of the actions of an object may involve sending messages to other objects. Variables and methods are defined via an inheritance hierarchy; they may be local to that object, defined for every object of a class, or even inherited from more abstract objects.

In our tutors, each subgoal of the curriculum is represented by an object, and higher-level objects act, in part, by asking their prerequisite objects to teach the needed prerequisite skills. We call the curriculum layer in such a system a curriculum-object lattice structure. The objects for a proposed tutor have the following content:

Declarative Knowledge

- Variables that identify how a given object's goals relate to the goals of other objects (i.e., which goals are prerequisite to the current one, and for which goals the current one is a prerequisite).
- Variables that identify how the knowledge an object is trying to teach relates to the knowledge other objects are trying to teach (pointers to the knowledge layer).
- Variables that represent the student's knowledge of the object's goal knowledge and functions that update those variables.

Procedural Knowledge

- Functions (methods) that generate instructional interventions based upon the student model held by the given object, including both manipulations of the microworld and various forms of coaching or advising.

- Functions that decide if the given object is to blame for problems that arise while objects for which it is prerequisite are in control (that is, if a student has trouble later on, the prerequisite objects can be asked if they see a reason for reviewing their lesson contents with the student).

A critical aspect of the diagnostic approach we are taking involves the notion of blame taking. The idea is to localize diagnosis. Existing approaches used in intelligent tutors use an *overlay* (Goldstein & Carr, 1977) approach. That is, they tend to take a sample of the student's performance and attempt to determine what sorts of deletions from the knowledge base would produce a system that behaved as the student does. This is very computationally intensive, and it depends on the assumption that the student's failings are all due to omissions in his knowledge; the student may also have misconceptions which cannot as easily be detected. To reduce the complexity of diagnostic processing in computer-based instructional systems, we have been developing a more localized view. At any given instant, a particular lesson is controlling the tutor; if the student has difficulties with that lesson, then control is transferred to the prerequisites for that lesson. If any of the prerequisite lesson objects finds reason to believe that what they taught was not adequately learned, it reteaches. This approach will be much more efficient than exhaustive, context-free diagnosis if most problems arise because an immediate prerequisite has not been learned adequately. However, it is likely to be more efficient any time that the prerequisite structure is adequate, i.e., that it captures the range of knowledge that could be missing. This is because the alternative is simply to search all of the knowledge space for an appropriate gap rather than to follow an optimizing search strategy.

We have described three layers for the intelligence that constitutes our proposed tutor: the aptitude layer, the curriculum layer, and the knowledge layer. The curriculum layer will be the driving layer of the system. At any given instant, a particular lesson object will control the course of processing. It will contain pointers to portions of the knowledge base and will report on its successes and failures to the appropriate object in the aptitude layer. In the course of responding to a report message from a lesson, the aptitude layer may take steps, such as adjusting risk-taking parameters, that involve changing the variables in various lesson objects. However, the basic controlling sequence will be driven by lessons that take control, ask for prerequisite lessons to be taught, integrate the prerequisite knowledge by presenting composite problems that involve multiple knowledge aspects all at once, and then notify the object that called them (for which they are prerequisite) that they are done. This lesson-driven approach sets the stage for eventual specification of a design approach that can be a replacement for the frame-oriented approach that has driven earlier generations of computer-assisted instruction.

SUMMARY: GENERAL DIMENSIONS FOR A COGNITIVE APPROACH TO THE MEASUREMENT OF ACHIEVEMENT

In what follows, we attempt to summarize ideas that could comprise a theoretical basis for the design of tests and assessment instruments to determine levels of knowledge and skill that are attained in the course of instruction. These ideas should be considered as a basis for test item construction coordinate with or prior to psychometric considerations.

Fundamentally, achievement measurement should be driven by the emerging cognitive theory of knowledge acquisition. We now realize that people who have learned the concepts and skills in a subject-matter domain have acquired a large collection of schematic knowledge structures. These structures enable understanding of the relationships inherent in their knowledge. We also know that someone who has learned to solve problems, to make inferences, and to be skillful in a subject-matter domain has acquired a set of cognitive procedures attached to knowledge structures that enable actions that influence learning, goal setting and planning.

At various stages of learning, there exist different integrations of knowledge, different degrees of procedural skill, differences in rapid access to memory and in representations of the tasks one is to perform. The fundamental character, then, of achievement measurement is based upon the assessment of growing knowledge structures and related cognitive processes and procedural skills that develop as a domain of proficiency is acquired. These different levels signal advancing expertise or passable blockages in the course of learning.

Achievement measurement theory, as we envision it, is at an early stage. Many of the ideas needed are yet to be worked out, but stimulating work has been done that gives indication of the shape of a guiding framework. Relatively speaking, we have most knowledge of differences between beginners and experts, but less knowledge of the intermediate stages and the nature of the transitions from level to level.

We can, however, on the basis of the work reported in this paper propose a tentative set of "dimensions" that comprise components of developing proficiency that might underlie the assessment of achievement. These dimensions are certainly covered to some extent in traditional forms of achievement assessment, but also may require new forms and methods of measurement. In any case, whether or not items take on new characteristics, they will be informed by a theoretical base which will drive more systematic rationales for interpretations of the meaning of test scores, particularly for diagnostic aspects necessary for instruction. We consider the following dimensions:

1. *Knowledge organization and structure.* As efficiency is attained in a domain, elements of knowledge and components become increasingly interconnected so that proficient individuals access coherent chunks of information rather than disconnected fragments. Beginners' knowledge is fragmentary, consisting of isolated definitions, and superficial understandings of the meanings of appropriate vocabulary. As proficiency develops, these items of information become structured, integrated with past organizations of knowledge so that they are retrieved from memory rapidly in larger units. The degree of fragmentation and structuredness and the degree of accessibility to interrelated chunks of knowledge becomes a dimension of assessment.

2. *Depth of problem representation.* It is now well known that novices recognize the surface features of a problem or task situation and more proficient individuals go beyond surface features and identify inferences or principles that subsume the surface structure. This growing ability for fast recognition of underlying principles is an indication of developing achievement and could be assessed by appropriate pattern recognition tasks in verbal and graphic situations. Certain forms of representation may be highly correlated wiht details of the ability to carry out a task or solve a problem. If this is the case, then test items might concentrate on assessing initial understanding and depth of representation and spend less time on the details of arriving at the correct answer.

3. *Quality of mental models.* People develop mental models of phenomena and situations with which they work. The nature of these representations is determined by what is useful for the tasks that need to be performed and the level of achievement that is required. One's mental model of a computer or a television set, of a mathematical proof, of an electric circuit, or the structure of DNA is dependent upon levels of knowledge and the processing requirements attached to performance. As tasks become more complex, these models are amended appropriately. There is a difference in the kind of knowledge required by the user, repairman, and designer of a television set. The nature of these models is an important dimension of achievement assessment; they indicate not only levels of task complexity that a person is capable of handling, but also the level at which the school requirements (and job demand) force people to think. The demands of school problem-solving tasks may require mental models less sophisticated than the curriculum implies. This discrepancy poses an interesting dilemma because when proficiency is assessed it is the model required by actual performance that is acquired and retained.

4. *Efficiency of procedures.* Carrying out procedures is an important aspect of many skills and is important also for effectiveness in higher level forms of problem solving and comprehension. Well-practiced procedures are significant for understanding and comprehension—for example, rephrasing or summarizing what one is reading is a performance characteristic of good comprehension; defining an audience and planning a structure are characteristic of good writing.

Such procedures need to be carefully assessed, but they cannot be measured in rote fashion. They must be assessed in terms of the effective goals that are guiding them. It is not enough to assess summarization and paraphrasing unless their effects on comprehension are considered. It is not enough to give an exercise in planning a composition, unless its effect on the writing process is engaged. The relationship between task understanding and efficient procedures is an important aspect of cognitive proficiency, and effective achievement measurement should exclude rote and piecemeal assessment of procedural skills that does not focus on performance goals.

5. *Automaticity to reduce attentional demands.* In investigations of competence, it has become evident that human ability to perform competing attention-demanding tasks is rather limited. When subtasks of a complex activity require simultaneous demands for attention, the efficiency of the overall task is affected. This fact has particular implications in the diagnostic assessment of the interaction between basic skills and advanced components of cognitive performance. As has been indicated, an example of this interaction has been of special interest in the investigation of reading and text comprehension, where attention may alternate between basic decoding skills of recognizing words and higher level skills of comprehension that integrate sentence ideas into memory. Although these component processes may work well when tested separately, they may not be efficient enough to work together. A slow, or inefficient, component process in interaction with other processes can lead to breakdowns in overall proficiency. If a task, such as reading, consists of an orchestration of basic skills and higher level strategic comprehension processes, then measurement procedures should be able to diagnose the inefficiencies in this complex performance.

The instructional implication is that in the development of higher levels of proficiency, basic skills should receive enough practice so that they become automatized and can be performed with little conscious attention. This leaves conscious processing capacity that can be devoted to higher level processes as necessary. A criterion for assessment then, is the level of efficiency or automaticity required for subprocesses to have minimal interference effects, i.e., whether the automaticity of a basic process has progressed to a point where it can facilitate and be integrated into the total performance of which it is a part. Has it reached a point so that further, more advanced learning and higher level performance can occur? Specific procedures for assessing automaticity might involve the measurement of response latency and of susceptibility to disrupting influences by simultaneous attention-demanding tasks.

6. *Proceduralized knowledge.* Modern learning theory has suggested that the course of acquisition of components of knowledge proceeds from an initial declarative form to compiled procedural form. In the early stage, we can know a principle or a rule or an item of specialized vocabulary without knowing the conditions under which that item of knowledge is applicable and is to be used the

most effectively. Studies of the difference between experts and novices indicate that beginners may have requisite knowledge but this knowledge is not bound to the conditions of its applicability. When knowledge is accessed by experts, it is always associated with indications of how and when it is to be used appropriately. The implication for measurement is that the progression from declarative to tuned procedural information is an indication of the development of achievement in an area of knowledge. Task analysis in various technical skills has shown that this progression can be assessed by qualitative differences in people's descriptions and definitions of their knowledge. Concepts, principles, and procedures can be measured in a way to determine the level of knowledge that is available to a learner. Test items can be comprised of two elements—information that needs to be known and conditions under which use of this information is appropriate. Our hypothesis is that advancing achievement will show changes in the level of knowledge from initial declarative knowledge to more complex combinations of actions and their conditions of use.

7. *Procedures for theory change.* As individuals learn, they solve problems and comprehend materials that foster further learning. This learning takes place on the basis of existing knowledge structures or theories held by students that can enhance or retard learning. With appropriate instruction, students can test, evaluate, and modify their current theories of knowledge on the basis of new information, and develop new schemata that facilitate more advanced thinking and problem solving.

While theories of knowledge held by students are a basis for new learning, current research has also emphasized that individuals hold naive theories, for example, at the beginning of a course in physics or economics, that make learning difficult. Even after a course of instruction, these naive theories persist, although students have learned, in some mechanical fashion, to solve problems in the course, but with little understanding. With this in mind, theories of knowledge become a target for assessment. The characteristics of a theory held by a student might indicate whether it is a tractable theory, amenable to change under certain instructional conditions, or whether the theory held is one that teachers find more intractable, that results in learning difficulties, and that requires additional instruction.

8. *Metacognitive skills for learning.* Metacognition is defined in a number of ways in the literature, but we consider here that aspect of it which refers to self-regulatory and self-management skills. Regulatory skills refer to generalized skills for approaching problems and for monitoring one's performance. These skills are called metacognitive because they are not specific performances or strategies involved in solving a particular problem or carrying out a particular procedure. Rather, they refer to the kind of knowledge that enables one to usefully reflect upon and control one's own performance. Representative kinds of regulatory performance include: knowing when or what one knows or does not know, predicting the correctness or outcome of one's performance, planning

ahead and efficiently apportioning one's time, and checking and monitoring the outcomes of one's solution or attempts to learn.

Research has indicated that these regulatory skills develop with maturity and that they may be less developed in students with learning disabilities or performance difficulties. It is likely that these skills appear in various forms and levels of competence over a wide range of individuals. An especially interesting characteristic of these skills is that they may be the particular aspect of performance that facilitates transfer to new situations. Individuals can be taught a rule or procedure that improves their task performance, but it is also important to learn how that rule is to be used and how to monitor its use. Self-regulatory activities of this kind are important candidates for assessment. Tests of an individual's competence in these metacognitive skills might be important predictors of success of the kind of problem-solving ability that results in learning.

Achievement testing as we have defined it is a method of indexing stages of competence through indicators of the level of development of knowledge, skill and cognitive process. These indicators display stages of performance that have been attained and on which further learning can proceed. They also show forms of error and misconceptions in knowledge that result in inefficient and incomplete knowledge and skill and that need instructional attention.

Achievement measurement defined in this way needs to be informed by theories of the acquisiton of subject-matter knowledge, by the development of knowledge and skill, and by various dimensions of performance such as degree of structure, automaticity, forms of representation and procedural efficiencies that indicate the growing and developing competence. We have speculated on possible indicators, but anticipate that theories of subject-matter acquisition will suggest both general indicators of competent performance, and also specific indicators dependent upon the nature of the knowledge and skill being assessed.

These theories require investigation and research, but work is proceeding rapidly. We anticipate that increasing sophistication in theory will be brought to achievement measurement, just as increasing sophistication in psychometric analyses has been brought to the design of tests after test items have been constructed. In essence, our paper is a signal for new orientations in achievement testing that will need to rely on the interrelationships between knowledge of learning and development, assessment of the indicators of growing competence, and their relevance to methods of instruction.

Finally, achievement measurement, as we have defined it, is an integral part of an instructional system. Teaching and testing are not separable events. Perhaps the term "learning assessment" better conveys our meaning than "achievement test," because the forms of measurement we envision provide information about the performance characteristics of levels of competence attained and about steps that can be taken to facilitate further learning.

ACKNOWLEDGMENT

Arlene Weiner provided substantial editorial assistance in preparing this chapter. Work we report was supported by the Office of Naval Research, the Air Force Human Resources Laboratory, and the National Institute of Education. None of these agencies has read or approved this document, and no endorsement of it by any of them should be inferred.

REFERENCES

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

Anderson, R. (1978). Schema-directed processes in language comprehension. In A. M. Lesgold, J. W. Pellegrino, S. D. Fokkema, & R. Glaser (Eds.), *Cognitive psychology and instruction* (pp. 67–82). New York: Plenum.

Bartholomae, D. (1980). The study of error. *College Composition and Communication, 31*(3), 253–269.

Benbassett, J., & Bachar-Bassan, E. (1984). A comparison of initial diagnostic hypotheses of medical students and internists. *Journal of Medical Education, 59,* 951–956.

Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). An instructional redesign of reading lessons: Effects on comprehension. *Reading Research Quarterly, 17,* 462–481.

Belmont, J. M., & Butterfield, E. C. (1977). The instructional approach to developmental cognitive research. In R. V. Kail, Jr., & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 437–481). New York: Wiley.

Bonar, J. (1985, June). *Bite-Sized Intelligent Tutoring* (Technical Report) Pittsburgh: Learning Research and Development Center, University of Pittsburgh.

Borkowski, J. G., Cavanaugh, J. C., & Reichart, G. J. (1978). Maintenance of children's rehearsal strategies: Effect of training and strategy form. *Journal of Experimental Child Psychology, 26,* 288–298.

Bransford, J. D., Delclos, V. R., Vye, N. J., Burns, M. S., & Hasselbring, T. S. (1986, February). *Improving the quality of assessment and instruction: Roles for dynamic assessment* (Working Paper 1). Nashville, TN: John F. Kennedy Center for Research on Education and Human Development, Peabody College, Vanderbilt University.

Brown, A. L. (1978). Knowing when, where and how to remember: A problem of metacognition. In R. Glaser (Ed.), *Advances in instructional psychology,* (Vol. 1, pp. 77–165). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell & E. M. Markman (Eds.), *Cognitive development* (Vol. 3 of P. H. Mussen (Ed.), *Handbook of child psychology* (pp. 77–166). New York: Wiley.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2,* 155–192.

Brown, A. L., & French, L. A. (1979). *The zone of potential development: Implications for intelligence testing in the year 2000* (Technical Report No. 128). Champaign-Urbana, IL: Center for the Study of Reading.

Brown, J. S., & Van Lehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science, 4,* 379–426.

Brown, J. S., & Van Lehn, K. (1982). Towards a generative theory of "bugs." In T. P. Carpenter, J. M. Moser, & T. A. Romberg (Eds.), *Addition and subtraction: A cognitive perspective* (pp. 117–135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Burton, R. R., & Brown, J. S. (1982). An investigation of computer coaching for informal learning activities, in D. Sleeman & J. S. Brown (Eds.), *Intelligent tutoring systems* (pp. 79–98). Orlando, FL: Academic Press.

Carey, S. (1985). Are children fundamentally different kinds of thinkers and learners than adults? In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills: Vol. 2. Research and open questions*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Champagne, A., Klopfer, L., & Anderson, J. H. (1980). Factors affecting the learning of classical mechanics. *American Journal of Physics, 48,* 1074–1079.

Chase, W. G., & Chi, M. T. H. (1981). Cognitive skill: Implications for spatial skill in large-scale environments. In J. Harvey (Ed.), *Cognition, social behavior, and the environment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152.

Curtis, M. E. (1980). Development of components of reading skill. *Journal of Educational Psychology, 72,* 656–669.

Curtis, M. E., & Glaser, R. (1983). Reading theory and the assessment of reading achievement. *Journal of Educational Measurement, 20,* 133–147.

deGroot, A. D. (1965). *Thought and choice in chess*. The Hague: Mouton.

deGroot, A. D. (1966). Perception and memory versus thought: Some old ideas and recent findings. In B. Kleinmuntz (Ed.), *Problem solving: Research, method, and theory*. New York: Wiley.

deKleer, J., & Brown, J. S. (1985). A qualitative physics based on confluences. *Artificial Intelligence, 24,* 7–84. Journal issue reprinted as D. G. Bobrow (Ed.). (1985). *Qualitative reasoning about physical systems* (pp. 205–280). Cambridge, MA: MIT press.

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.), *Categories of human learning*. Orlando: FL: Academic Press.

Gentner, D., & Gentner, D. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens (Eds.), *Mental models*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gitomer, D. (1984). *A cognitive analysis of a complex troubleshooting task*. Unpublished dissertation, University of Pittsburgh.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist, 39,* 93–104.

Goldstein, I., & Carr, B. (1977, October). The computer as coach: An athletic paradigm for intellectual education. *Proceedings of 1977 Annual Conference, Association for Computing Machinery,* Seattle, pp. 227–233.

Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. In L. N. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoepfner, R. (1978). Achievement test selection for program evaluation. In M. J. Wargo & D. R. Green (Eds.), *Achievement testing of disadvantaged and minority students for educational program evaluation*. Monterey, CA: CTB/McGraw-Hill.

Hull, C. L. (1943). *Principles of behavior*. New York: Appleton-Century-Crofts.

Hull, G., Ball, C., Fox, J., Levin, L., & McCutchen, D. (1985). *Computer detection of errors in natural language texts: Some research on pattern matching*. Paper presented to the American Educational Research Association, Chicago.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Larkin, J. H. (1983). Teaching problem solving in physics: The psychological laboratory and the practical classroom. In D. T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208,* 1335–1342.

Lesgold, A. M. (1984a). Acquiring expertise. In J. R. Anderson & S. M. Kosslyn (Eds.), *Tutorials in learning and memory: Essays in honor of Gordon Bower*. San Francisco: W. H. Freeman.

Lesgold, A. M. (1984b). Human skill in a computerized society: Complex skills and their acquisition [Presidential address to the Society for Computers in Psychology]. *Behavioral Research Methods, Instruments & Computers, 16,* 79–87.

Lesgold, A. M., & Curtis, M. E. (1981). Learning to read words efficiently. In A. M. Lesgold & C. A. Perfetti (Eds.), *Interactive processes in reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lesgold, A. M., & Perfetti, C. A. (1978). Interactive processes in reading comprehension. *Discourse Processes, 1,* 323–336.

Lesgold, A. M., Resnick, L. B., & Hammond, K. (1985). Learning to read: A longitudinal study of word skill development in two curricula. In T. G. Waller & G. E. MacKinnon (Eds.), *Reading research: Advances in theory and practice* (Vol. 4, pp. 107–138). Orlando, FL: Academic Press.

Lesgold, A. M., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (in press). Expertise in a complex skill: Diagnosing X-ray pictures. In M. T. H. Chi, R. Glaser, and M. Farr (Eds.), *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lidz, C. S. (Ed.). (in press). *Dynamic assessment: Foundations and fundamentals*. New York: Guilford Press.

Logan, D., & Eastman, R. (1986). *Mental models of electronics troubleshooting*. Paper presented to the annual meeting of the American Educational Research Association, San Francisco.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science, 210,* 1139–1141.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7,* 44–64.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Perfetti, C. A., & Hogaboam, T. W. (1975). The relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology, 67,* 461–469.

Perfetti, C. A., & Lesgold, A. M. (1977). Discourse processing and sources of individual differences. In P. Carpenter & M. Just (Eds.), *Cognitive processes in comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Perfetti, C. A., & Lesgold, A. M. (1979). Coding and comprehension in skilled reading. In L. B. Resnick & P. Weaver (Eds.), *Theory and practice of early reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Qualitative Reasoning about Physical Systems. (1984). [Special issue]. *Artificial Intelligence,* 24 (1–3). [Also published as D. G. Bobrow (Ed.). (1985). *Qualitative Reasoning about Physical Systems*. Cambridge, MA: MIT Press.]

Riley, M. S. (1985). *Structural understanding in performance and learning*. Unpublished doctoral dissertation. Pittsburgh, PA: University of Pittsburgh.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153–196). Orlando, FL: Academic Press.

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. G. Brown & A. Collins (Eds.), *Representation and understanding: Studies in cognitive science*. Orlando, FL: Academic Press.

Rumelhart, D. E. (1981). *Understanding understanding*. La Jolla: University of California, Center for Human Information Processing.

Schneider, W. (1984). Practice, attention, and the processing system. *Behavioral and Brain Science, 7,* 80–81.
Shaughnessy, M. (1977). *Errors and expectations.* New York: Oxford University Press.
Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 4,* 481–520.
Siegler, R. S. (1978). The origins of scientific reasoning. In R. S. Siegler (Ed.), *Children's thinking: What develops?* Hillsdale, NJ: Lawrence Erlbaum Associates.
Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. S. Siegler (Ed.), Children's thinking: What develops? Hillsdale, NJ: Lawrence Erlbaum Associates.
Simon, H. A., & Chase, W. G. (1973). Skill in chess. *American Scientist, 61,* 394–403.
Spence, K. W. (1956). *Behavior theory and conditioning.* New Haven, CT: Yale University Press.
Van Lehn, K. (1983a). Human procedural skill acquisition: Theory model and psychological validation. *Proceedings of the National Conference on Artificial Intelligence,* Washington, DC.
Van Lehn, K. (1983b). On the representation of procedures in repair theory. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 201–252). Orlando, FL: Academic Press.
Van Lehn, K. (1983c). *Felicity conditions for human skill acquisition: Validating an AI-based theory* (Report CIS-21). Cognitive and Instructional Sciences Series, Xerox Palo Alto Research Center: Palo Alto, CA.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* (M. Cole, V. John-Steiner, & E. Soubermen, Eds. & Trans.). Cambridge, MA: Harvard University Press.