DOCUMENT RESUME

ED 384 666 TM 023 973

AUTHOR Snow, Richard E.; Mandinach, Ellen B.
TITLE Integrating Assessment and Instruction: A Research and Development Agenda.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-91-8
PUB DATE Feb 91
NOTE 192p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC08 Plus Postage.
DESCRIPTORS Agenda Setting; Diagnostic Tests; *Educational Assessment; Educational Objectives; Instructional Effectiveness; *Integrated Activities; Literature Reviews; *Needs Assessment; *Psychometrics; *Research and Development; Teacher Evaluation
IDENTIFIERS *Performance Based Evaluation

ABSTRACT

The long-range research and development needs for integrating assessment into instruction are discussed, and how ready the educational community is to begin making blueprints for this integration is explored. This paper addresses the problems of learning progress, diagnostic assessment, the design and implementation of performance tasks, and mapping collections of tasks in order to make an instructional domain to guide instructional adaptation. The paper is a formative and conceptual discussion, rather than a comprehensive literature review. It is, in effect, a review of issues rather than one of accumulated evidence. The following basic issue categories are explored: (1) instructional goals, domains, and treatments; (2) reference tasks and teacher assessments; (3) the nature of learning from instruction; (4) diagnostic assessment for instructional use; and (5) psychometric problems. A summary of the research agenda and recommendations is included. Building domain topographies for instructional assessment is a step toward theories of understanding for the domains addressed. Three tables and three figures illustrate the discussion. (Contains 296 references.) (SLD)

RESEARCH REPORT

INTEGRATING ASSESSMENT AND INSTRUCTION: A RESEARCH AND DEVELOPMENT AGENDA

Richard E. Snow
Ellen B. Mandinach

Educational Testing Service
Princeton, New Jersey
February 1991

INTEGRATING ASSESSMENT AND INSTRUCTION:

A RESEARCH AND DEVELOPMENT AGENDA

Richard E. Snow
Stanford University

Ellen B. Mandinach
Educational Testing Service

Running head: INTEGRATING ASSESSMENT AND INSTRUCTION

Acknowledgement

This paper was commissioned by the Educational Testing Service as one way to help inform its research programs aimed at improving assessment in the service of learning. A companion document presents a more complete bibliography for use as a reading list by ETS researchers (see McVey, Snow, & Mandinach, 1989). The authors thank Mary McVey for her substantial efforts in organizing the literature on which this paper is based. In addition to many helpful suggestions and criticisms provided by ETS reviewers, the authors wish to acknowledge also the critical reviews and suggested revisions of the paper provided by Richard Burton, Lyn Corno, Hugh Cline, Lee Cronbach, Larry Frase, Drew Gitomer, James Greeno, Edward Haertel, Berner Lindstrom, Denis Phillips, Lee Shulman, and Marshall Smith. Responsibility for the substance of the document and any errors contained herein, however, remains with the authors. The authors also wish to thank Ruby Brice for her patience and diligence in working on various drafts of the manuscript.

Table of Contents

INTEGRATING ASSESSMENT AND INSTRUCTION

Introduction

If assessment and instruction can be integrated, education will improve. Most educators and measurement specialists believe that educational tests will be more useful if they are connected more closely to day-to-day instructional needs and goals. They also believe that teachers will make better, more direct use of systematic assessments when those assessments are embedded within instruction. Teachers have long sought this sort of interstitial help from home-made instruments. Textbooks include intermediate discussion questions and exercises with the same intent. However, these practices have been unsystematic, with little guidance from research; experience with them has not been cumulative.

There have been some research-based attempts to redesign tests, embed them in instructional sequences, and make subsequent instruction contingent on learner response. The movements toward behavioral objectives, criterion- or domain- or content-referenced tests, and some adaptive instructional designs such as individually prescribed instruction or mastery learning, all exemplify attempts to connect assessment and instruction. But in all these efforts, the instructional tasks and the tests remain distinct; instructional design and test design are independent activities. In contrast, one can imagine a unification of instructional and assessment design in which both learning and its assessment derive from the same tasks and activities. This would seem to require coordinated theories of learning progress and of diagnostic assessment for adaptive teaching in an instructional domain. Theories of this sort are not now available.

It can be argued that integrating assessment and instruction is an engineering design problem--one has to invent an airplane before theory and research can be formulated to reach an understanding of aerodynamics. But one never really knows when early attempts at invention have proceeded far enough profitably to set a research agenda. And it may be that long-range views of research needs can help focus short-range decisions on design and development, at least of prototypes. We believe the time is now ripe to set an agenda for integrating assessment and instruction. Below, we present one view of the long-range research and development needs for integrating assessment in instruction, and test how ready we are to begin making blueprints.

Overview

Advances in cognitive, developmental, differential, and instructional psychology have recently become substantial enough to justify serious attention to the possibility of integrated instructional and assessment systems. New ways of conceptualizing and studying such systems are available. New technologies now make such systems more practicable. To capitalize on these scientific and technological advances, however, a substantial transformation of test research and development practice will be required. Test designs based only on classified instructional objectives and psychometric analysis of items associated with those objectives will not suffice. But the new substantive psychological theories also do not suffice. No one yet has principles for building integrated assessment-instructional systems that will be valid and useful for teachers.

Consider some of the assessment needs in everyday instruction. In expert human tutoring, teaching and testing functions are so meshed that the line between the two is usually invisible to the learner. Often the same task or dialogue about it serves both

functions. And teaching is adapted not only to the results of each learning step but to a developing understanding of the individual as a learner. By analogy, computerized instruction is deemed potentially "intelligent" not just because each instructional step can be conditioned on the previous learner response, but because both instruction and assessment can be made adaptive to a deeper, continuing diagnosis of learner progress; the two functions should work interactively across sequences of instruction to build both progress and diagnosis. Unfortunately, of course, the teaching and testing functions become distinct and disconnected as optimal tutoring is transformed by practical necessity into full classroom management. But even here, one can imagine an instructional assessment system running separately, using computerized exercises or other organized learning activities, that remains in close coordination with teacher classroom instruction. The ongoing diagnosis of learning progress and the adaptation of instruction to it remains the teacher's role, but the supporting system is explicitly designed to serve this role. Whether teacher-based or computer-based or both, intelligent teaching is seen to rest on closely articulated instructional and assessment functions that guide adaptation to individual learner progress. To improve everyday instruction, we need to understand this articulation.

We are not yet close to this goal. But several programs of research and development have now begun to address it. Substantial advance in such programs, however, depends on having in the early stages a clear conception of several difficult problems to be faced and some approaches to studying and overcoming them. These problems can be introduced here as a set of four general questions, though there are many elaborations to be taken up later. The questions are:

1. What constitutes learning progress toward mastery in an instructional domain?

2. What constitutes diagnostic assessment of this learning progress for instructional use?
3. How might performance tasks that provide such assessment be designed and evaluated?
4. How might collections of such tasks be mapped into an instructional domain to guide instructional adaptation?

The purpose of this paper is to address these problems, to suggest possible approaches to their solution, and thereby to assist later system development and evaluation. It uses relevant literature where possible, and identifies needs for further basic research, as well as research aimed at system design. But the paper is intended mainly as a formative conceptual step, not a comprehensive literature review or a specific research and development proposal. In effect, we are reviewing research and development issues in the literature rather than accumulated evidence from that literature.

Recent Developments And Example Systems

New research in instructional psychology, psychometrics, and also in computer technology, should inform the conception, design, and evaluation of any attempt at integrating instruction and assessment. There are at present a number of concrete attempts at system design to accomplish this goal, in whole or part. Lessons can be learned from examining these developments. Current projects may benefit from contracts with other projects. For example, there may be exchangeable parts from one effort that can improve another.

In this section we review four prominent example programs. The first two represent designs based on current cognitive and developmental theory in relation to instruction; both

attempt to provide more direct assessment of learning in particular instructional tasks that teachers might find troublesome on a daily or weekly basis. The second two come more from advances in computer technology that attempt to capitalize on present cognitive and psychometric theory for instructional and assessment purposes; these designs aim more at the level of whole courses of instruction.

Notes on Recent History

It is important first to recognize three recent transitions in the psychology of instruction, the psychology of educational measurement, and the development of instructional technology. The three did not occur in synchrony, and they are only now beginning to connect. Together they make a useful preface for our examples, but also for the consideration of other possible designs.

From behavioral to cognitive analysis in instructional design. Through research on training born of World War II and on programmed instruction during the 1960's, it became clear that instructional and assessment technologies based only on behaviorist principles were quite limited (Gagné, 1965; Glaser, 1963). Case and Bereiter (1984) have reviewed this argument; we abstract their view rather than others because their dissatisfactions led to the development of our first example program.

Behaviorist language is not sufficient for specifying objectives for school learning, because it degrades the definition of cognitive concepts and procedures to their behavioral signs (i.e., their recitation or application; rote recitation is not excluded). It also leaves out aspects of understanding demonstrably crucial to later attainments. Reinforcement principles, while sometimes useful in instruction, are neither necessary nor sufficient to advance learning. Furthermore, test items matched one-to-one with behavioral objectives

misdirect attention toward the test and away from the deeper instructional objectives for which the test task is intended to be an indicator. Most important, behaviorist methods provide no way to identify what instructional steps lead toward the desired objectives unless these objectives are susceptible to successive approximation conditioning. Behavioral approaches have had some success, notably in special education (see Deno, 1985; Fuchs & Fuchs, 1990). But these approaches do not escape the above criticisms, and they concentrate on observation of learners during fixed instruction, not on assessment-instruction interaction and adaptation.

Case and Bereiter (1984) show how the shift from behaviorist to "first wave" cognitivist instructional technology solved these problems but met others. Gagné's (1965, 1968) hierarchical analysis of instructional objectives is the principal case. Successive analyses of subgoals prerequisite to end-goals produces a hierarchical, branching network to identify both instructional and assessment steps in sequence. The network then is used to design instructional sequences to promote the succession of attainments up the hierarchy.

Gagné's approach has also had some successes. Hierarchical task analysis seems most applicable to the definition and assessment of component intellectual skills and strategies in well-structured domains such as elementary arithmetic, and remains useful as a suggestive technique for initial structuring of our understanding in many other domains. However, the research evidence shows that it is also neither necessary nor sufficient for specifying instructional objectives, designs, and assessments, even in arithmetic. The approach has occasionally been found unsuccessful with young children and adults who appeared to possess the same prerequisite skills as intermediate students for whom the instruction had worked. Problems also crop up when the material to be learned poses

inherent difficulties that tax the limits of human information processing at a particular stage of development. Transfer to different situations may also be limited. Finally, not all intermediate steps in a hierarchy appear to be necessary. Some learners seem to skip some instructional steps, later filling in the unpracticed subskills in top-down fashion without explicit instruction. Some learners invent strategies they have not been taught and so advance in unspecified or unexpected ways (Resnick, 1976).

Gagné's example nonetheless encouraged effort toward a technology of instructional design; and many elaborations, variations, and related ideas have evolved (see Gagné, 1987; Gagné, Briggs, & Wager, 1988; Reigeluth, 1983, 1987a; Snelbecker, 1985). The two volumes edited by Reigeluth (1983, 1987a) collect several representative examples and show their application to a specific lesson. In Scandura's (1983; Stevens & Scandura, 1987) Structural Learning Theory, instructional steps and sequences are specified using paths through rules to be mastered. Instruction is focused on paths or rule components not yet mastered, as identified by test items associated with each. The items are derived directly from the rule objectives and instances; they usually require the demonstration of a procedural path for the rule. The tests are separate from the instruction, however. Another example is Merrill's (1983, 1987) Component Display Theory, in which instructional objectives are defined according to content and level of performance required in narrowly specified domains. Prescriptive rules connect a taxonomy of alternative presentation methods to objectives in the content by level matrix. The learner is given some control over the sequence of instructional steps chosen. Test items can be connected to objectives and presumably provide feedback loops to learner and system. Beyond this micro-instructional focus, Reigeluth's (1987b; Reigeluth & Stein, 1983) Elaboration Theory

applies similar procedures to more macro-level instructional steps and decisions. Along with its predecessors, it uses test items that reflect behavioral or procedural specifications of objectives, immediate feedback, and organizes instruction and testing in simple-to-complex hierarchical form. It also uses low-risk self-tests.

Formulations such as those cited offer prescriptive theories for instructional designs with various characteristics, and corresponding strengths and weaknesses. It is perhaps too early to judge their value. In some instructional settings, one or another may serve well. In our view, however, all of these approaches move too quickly to prescription without coming fully to grips with the psychology of instructional variables and performance, and of diagnostic assessment for instructional adaptation. None truly attempts to integrate instruction and assessment. As Baker and O'Neil (1987) note, assessment considerations are usually subordinate in this work. At the least, they argue, continuing research in this line needs to build up an assessment database which might support inferences about cognitive and attitudinal development during such instruction, and thus suggestions for more directly integrated assessment devices.

From content to construct validation in test design. Through the same period, dissatisfaction with educational tests also arose, and from some of the same sources (e.g., Glaser, 1963). It was argued that conventional, norm-referenced tests did little to improve instruction. Behaviorist language was also found inadequate for non-behaviorist test specifications for reasons already noted (see also Haertel & Calfee, 1983).

The first consequences were test designs variously called criterion-, or content-, or domain-referenced, as opposed to norm-referenced measures. The aim was to connect test items specifically to the instructional objectives to be taught and thus to reflect changes

due to learning from instruction more directly. Items would be selected on this criterion, not on the basis of correlational statistics. As previously noted, these tests were not integrations of assessment and instruction; they could be connected to specific objectives, but they stood apart from instruction toward these objectives. Furthermore, it soon became apparent that it is the test interpretation that is "referenced", not the test design (Cronbach, 1984, 1989); many standardized achievement tests now also provide content-referenced interpretations. And tests purportedly designed for different interpretive uses have usually turned out to be substantially correlated.

It also became apparent that the new test designs did not solve problems of learning assessment for instructional purposes. They could be closely matched to local instructional objectives only if content domains and goals were carefully and elaborately prespecified in stimulus-response detail (see Popham, 1975). But this narrowed the notion of what is learned, leaving out divergent, deeper, and longer range developments and transfer objectives (Cronbach, 1982a, 1982b). Furthermore, this specification relied on rational taxonomies or content-by-process tables of objectives of the same sort as did tests used for norm-referenced interpretations. Whatever the purpose of the test, items are written to fit these tables of specifications, and the test is judged only by the adequacy of its sampling of the domain as defined. It may be useful to offer interpretations that translate test scores into content- or objective-based statements. But listing the bits and pieces of what learners can and cannot do in each portion of a rationally defined domain, at some point in time, is hardly a diagnosis that a teacher can use to guide further instructional steps for particular learners toward deeper, long-range learning goals of transferable understanding and skill.

A chief obstacle to developing more adequate learning assessments, in our view,

has been the failure to recognize that logical taxonomies and content specifications do not represent the psychological structures and processes that underlie performance in educational domains. Indeed, they tend to suppress, misrepresent, or neglect what the performer does mentally. The next wave of theory and research on educational assessment will come from recognition that interpretations of achievement measures rest on psychological constructs and have to be validated as such. Construct validity is the essence of all test validation enterprises (Cronbach, 1980, 1988a, 1988b; Messick, 1989); the traditional concepts and methods of content validation no longer justify most uses of educational measures. Developments in cognitive and conative psychology suggest new ways of thinking about the constructs underlying educational measures, new constructs to be assessed, and new kinds of test designs (Embretson, 1985; Glaser, Lesgold, & Lajoie, 1987; Nitko, 1989; Snow, 1989c; Snow & Lohman, 1989). Psychometric theory is responding with reexaminations of its basic models, and expansions to address some of the new issues in test design and validation. A central focus in this movement is diagnostic assessment of learning progress (Embretson, 1987, 1990; Frederiksen, Glaser, Lesgold, & Shafto, 1990; Frederiksen, Mislevy, & Bejar, in press).

Diagnosis involves more than assessing the learner's mastery of some instructional content domain. There are understandings and misunderstandings, abilities and inabilities, and attitudes that learners bring to or develop during instruction; assessing these aptitude differences for further learning is also part of diagnosis. Transfer assessment is part of diagnosis. And learning, transfer, and individual differences therein, have their conative and affective, as well as cognitive aspects. As with the newest developments in instructional design, it is too early to judge which initiatives in assessment design will bear fruit. Our

purpose here is to help bring at least some of these developments into alignment by considering concrete cases. It should be helpful also to have a picture of parallel developments in instructional and measurement technology.

Computerized measurement and instruction. Advances in technology are likely to facilitate the integration of instruction and assessment, since they help shift the focus of assessment from products to processes and provide access to and recording of interactive instructional and assessment activities. Both features promise to provide more diagnostically useful information for both student and teacher.

Bunderson, Inouye, and Olsen (1989) review the development of relevant technology, distinguishing four generations of computer-based techniques. The first two were focused on outcome measurement for institutional purposes; the last two have aimed at process assessment primarily to serve individuals.

The first generation, computerized testing, merely automated the procedures of traditional measurement. Existing instruments were translated for computerized administration, making possible greater standardization, additional performance measures (e.g., response latencies), and increased efficiency in scoring, reporting, and interpreting results. The second generation, computerized adaptive testing, provided item selection determined by examinee performance on previous items. This facility coupled with item response theory (Lord, 1980), permitted equally accurate measurement at all levels of ability and increased efficiency in test administration individualized for each examinee.

The third generation, continuous measurement, now focuses on learning progress assessment. An examinee is located within a learning space by assessing performance on tasks calibrated for difficulty as learning milestones leading to specific outcomes. The

milestones can be embedded unobtrusively within a curriculum. The aim is to provide continuous measurement linked to instruction. No such systems have been completed or fully evaluated to date, but several implement the basic principles. The TICCIT system (Bunderson, 1973) is perhaps the best example.

The fourth generation aims at "intelligent" measurement--the application of knowledge-based computing to complex responses, to generate more sophisticated interpretations and advise on remediation. Research on intelligent tutoring should make important contributions to the development of such systems. As Bunderson and colleagues note, the intersection of intelligent tutoring and measurement is new (but see Frederiksen & White, 1990). Tutor design has not considered psychometric issues. Similarly, measurement research has not provided much help for research on tutoring. The intersection is critical if "intelligence" is indeed to be built into computerized instructional systems.

A comparable progression of technological advances can be traced for computer-assisted instruction (see Barr & Feigenbaum, 1982). A first application was computer-based drill-and-practice, focused on narrowly specified learning outcomes and used primarily to supplement classroom instruction. The skills targeted were low-level and content-specific. Instruction took the form of direct practice with simple correct/incorrect feedback, used primarily for remedial purposes. A second approach has used commercial educational games and simulations as instructional tools. Although possibly addressing some higher-order cognitive skills, these products have not had explicit ties to school curricula or instructional objectives and their design has often violated sound principles of learning and motivational theory (see Lepper & Chabay, 1985; Lepper & Malone, 1987;

Malone & Lepper, 1987). Without clear instructional links, they become discovery learning problems; many students are unable to benefit from the implicit instruction (Mandinach, 1984, 1987).

The third generation sought to provide problem solving instruction, subsumed within the context of learning computer programming. The best example is LOGO. The fourth generation, as identified above, is intelligent tutoring in which the full capability of artificial expert systems is used. Although the first attempts suffered from limited conceptions of learner cognition, instruction, and diagnosis, more recent attempts have been built on more substantial cognitive theories (see Wenger, 1987).

From all of the above starting points, current work strives toward "deeper" psychological theories of integrated instruction and assessment. We turn now to a review of four example systems that attempt to do this. Although, two are designed mainly for human teacher use and two rely heavily on computer technology, each might be adapted to either form or some mixed form.

Developmental Assessment and Learning Strategies

Some new research emphasizes cognitive developmental analysis to identify and improve learning strategies, exemplified here by the work of Case (1985) and his colleagues (Case & Bereiter, 1984; Case & Sandieson, 1988; Case, Sandieson, & Dennis, 1986), but also Siegler (1986) and his colleagues (Siegler & Campbell, 1989, 1990). Some of this work seeks to create computer simulation models to help specify the internal cognitive structures and processes that may lead to observable features of performance, exemplified also by the work of Greeno (1986a, 1988). The approach uses analyses of expert performance in the relevant instructional domain as an important guide (see also Chi, Glaser, & Farr, 1988).

Basic concepts and procedures. The developmental approach devised by Case and Bereiter (1984) follows these basic steps:

1. Identify a task that reflects the educational objective to be taught and develop a measure for assessing student success or failure on this task.
2. Diagnose the strategy experts use for succeeding at this task.
3. Diagnose the strategy younger or less experienced students use when they succeed and also when they fail at this task.
4. Design an instructional sequence for showing students why their current strategy is inadequate, and for enabling them to assemble a more powerful strategy, to bring them from one level to the next in the course of instruction.
5. Minimize the working memory load of the instructional program by simplifying the structure of the successful strategy and/or by identifying a sequence of instructional steps from the unsuccessful to the successful strategy so that the number of novel cues or cognitive operations at each step is reduced to a minimum.
6. Provide sufficient practice at each step so that the operations to be applied become automatic, thus further reducing working memory load.
7. Move on to the next step only after performance becomes automatic.

It is important to recognize that the task chosen in Step 1 embodies the overall instructional objective; the task and the measure are intrinsic to the goal, not extrinsic as a separate set of items merely associated with it. Steps 2, 3, 4, 5 then acknowledge not only the cognitive prerequisites to task performance emphasized by Gagné and other instructional designers, but also the cognitive developmental aspects of learning. Steps 2 and 3 imply that there may be "natural" organizations and progressions in cognitive strategy from younger or less experienced students to older and more expert students. Step 4 implies that instruction might in some sense recapitulate this development. Step 5 attempts to simplify instruction in cognitive rather than simply behavioral terms. In particular, instruction is designed to fit within the load or capacity limits of the learner's working memory. Working memory capacity is a fundamental aptitude difference among learners in Case's (1985) theory.

Through a series of empirical examples, Case and Bereiter (1984) demonstrate how such an instructional design can work in elementary arithmetic, reading, and writing. Examples are also available in aspects of social studies, physical education, and life skills. In all these studies, it seems clear that complex performance involves flexible use of a variety of strategies, but that a definite developmental hierarchy of these strategies can be defined. Learning difficulties typically arise from use of low level strategies. Instruction of the sort described, when focused on particular strategies, has been shown to move learners toward more effective higher level strategies. The research also shows the importance of understanding the learner's mental representation of goals for particular tasks. A strategy then is defined as a means of bridging the gap between two such representations--one for the present task situation and one for the goal. In contrast to

learning hierarchy approaches, this developmental approach uses a global representation of the ultimate instructional goal from the start, rather than focusing on the progression of subgoals and introducing the final goal only at the end.

An example. In later research, Case et al. (1986) have given a detailed example demonstrating how the approach works in remedial teaching of time telling with analog clocks. They also describe two variations on the procedure, one called strategy incrementation, the other conceptual recapitulation. The two methods follow the same steps, as listed above. They differ, however, in their relative emphasis on using whatever strategies will lead toward a particular task criterion versus identifying the deeper conceptual structures and understandings that underlie a broad array of strategies and goals and their naturally occurring developmental sequences.

An expert strategy consisting of four cognitive processing phases was first hypothesized and validated with a sample of adults performing time-telling tasks based on a model of an analog clock. It was noted that adults all seem to use the same four phases, though they differ in strategy (e.g., they sequence the phases differently). The phases are: (a) a global scan to locate hour and minute hands, and identify familiar configurations; (b) determination of the hour value by detecting a close-to-hour pattern or, if no such pattern is clear, storing the observed hour value in working memory; (c) determination of the minute value to the nearest multiple of five, distinguishing on the mark, minutes to, and minutes past, on right and left sides of the clock face; and (d) determination of the minute value to the nearest single unit, adapting response with incremental or decremental strategies applied to results of phase three. Children's performance in each phase was assessed using similar tasks and compared to adult performance to identify strategic

differences in detail.

The strategy incrementation instruction was then designed to focus on each child-adult difference in turn, beginning with the simplest and ending with the most complex of these differences, always minimizing demand on working memory. The conceptual recapitulation instruction also included similar strategy exercises but was focused more on understanding; it introduced the conceptual as well as the procedural knowledge children would likely acquire in each normal level of cognitive development. Based on Case's (1985; see also Siegler, 1976, 1978) theory of normal intellectual development, these levels are: (a) a predimensional stage (during ages 3 1/2 to 5), in which simple polar variables are understood; (b) a unidimensional stage (during ages 5 to 7) in which the action of polar variables is combined with the concept of number to understand the operation of continuous quantitative variables; (c) a bidimensional stage (during ages 7 to 9) in which two continuous variables can be combined; and (d) an elaborated and integrated bidimensional stage (during ages 9 to 11) in which trade-offs and third variables are understood to affect the two-variable system.

It is important to note that both instructional exercises and the embedded assessments were based on the same clock models and related time-line depictions. The assessment tests typically reflect the four level developmental theory for time-telling as well as for balance beam problems and other similar conceptual domains. The recapitulation instruction includes much of the exercise with clock details used in the incremental approach, but it focusses centrally on developing the conceptual understanding into which these details fit. Both approaches were shown to produce substantial improvement from pretest to posttest, but the developmental recapitulation instruction was

by far the more powerful on transfer items. It also evidenced no backsliding between training and posttest. In other work, the recapitulation treatment has shown continuing improvement to delayed posttests given weeks later. Case and his coworkers emphasize that the power of this treatment lies in its diagnosis and instruction of both conceptual understanding and performance strategies, as well as its emphasis on the normal progression of such development. Although the Case theory posits a common sequence of cognitive structure development across different tasks and a common form of structure across tasks within an age group, the recapitulation treatment does not depend on the validity of this theory. As long as some sort of conceptual progression can be found to underlie strategic progression in an instructional domain, a recapitulation curriculum can be designed using this progression as a "route map".

Case and Sandieson (1988) have generalized the approach to help identify and develop the central conceptual structures that seem to underlie transitions to higher levels of thinking and reasoning in elementary mathematics and science. They show that the quantitative concepts and skills developed at one stage serve as tools of thought in progressing through subsequent stages, and that training that uses the conceptual recapitulation approach in quantitative reasoning transfers to other areas, such as scientific and social reasoning, time telling, and money handling. They also outline a view of mathematics and science instruction, and the integration of these two curriculum fields, based on their developmental theory.

Related research. Whereas Case and his colleagues have sought principles of instructional design based on a general developmental theory, some other developmental approaches have focussed more on the domain specific knowledge and skills needed in

particular subject-matter areas. General cognitive theory remains a starting point, but the aim has been to examine the discontinuities as well as the continuities in the development of children's thinking in various domains, and particularly in science. The work connects with studies of quantitative and qualitative changes in knowledge and skill associated with the acquisition of expertise, often called the "novice-expert shift". Some investigators also see individual developmental changes as recapitulating theoretical developments in the history of science. A child's naive theory of the physical world, for example, is seen as resembling a medieval view of physics and then is substantially restructured to resemble a more modern scientific, or more adult, conceptual framework. Strauss (1988) collected examples of this research. Vosniadou and Brewer (1987) have provided a review that addresses instructional design implications.

Developmental knowledge restructuring of this sort is seen as resulting from continuing accumulation and elaboration of related knowledge in a schema, but also from the radical restructuring of schemas that is required to deal with anomalies; the parallel is with Kuhn's (1970) account of paradigm shifts in the history of science. Based on their own research on children's astronomical concepts, as well as their review, Vosniadou and Brewer (1987) suggest that radical restructuring is most likely brought about by instruction that takes the form of Socratic dialogue, in which a teacher helps the learner recognize anomalies and then guides the learner's reconstruction to eliminate misconceptions and produce a more comprehensive schema. They also emphasize the important role that analogies, metaphors and physical models can play in this process. Whether or not the historical parallel should be used in instruction, however, remains an important, but open question. Some researchers (e.g., McCloskey & Kargon, 1988) suggest that teaching in

science might use the historical theory transitions to help students identify and clear away their own naive theories. These and other writers (e.g., Carey, 1988; Wiser, 1988) do not go on to imply that science instruction should actually recapitulate the historical stages by teaching the early theories, but that is an implication to ponder in relation to Case's (1985) developmental theory and his conceptual recapitulation instruction.

The diagnostic assessment side of this work typically relies on simple performance tasks involving some scientific principle about which the learner gives predictions and explanations, coupled with one-on-one interviewing designed to draw out conceptions in the learner's own words. Similar methods have been used in the phenomenographic approach of Marton (1981, 1983; see also Marton, Hounsell, & Entwistle, 1984). The so called "weak" form of restructuring is indicated by expert-novice differences in simple misconceptions, in incomplete or misleading conceptions or strategies, or the absence of relevant conceptions among novices, in the perception of similarities or dissimilarities among elements in the domain, and in information processing during problem solving (see Chi, Glaser, & Rees, 1982). More radical, or "strong" restructuring is evidenced when new and old schema are seen to differ qualitatively in their individual concepts, in the structures connecting them, and in the domains they explain. Carey (1988) argues that the strong form of restructuring is a central feature of child cognitive development. If so, then detecting this sort of theory-change in a domain should be a central focus of assessment for instruction. Whether or not Carey's strong restructuring view is seen as inconsistent with Case's stage-wise recapitulation view, the two share the emphasis of identifying the states of central conceptual structures in a domain at different levels of development. Carey's weak form of restructuring, as in the novice-expert shift, is akin to Case's strategy

incrementation procedure.

A further point for assessment connects to other research on learner strategies. Siegler and Campbell (1989, 1990) have demonstrated that even young children acquire and use a diversity of strategies in solving particular problems of the sort met in elementary arithmetic, reading, time-telling, and balance beam tasks. Learners not only use different strategies for different problems but also shift strategies within a problem as a function of the difficulty characteristics of the task. These findings are consistent with other demonstrations of multiple strategies and strategy shifting in high school and college students (Kyllonen, Lohman, & Woltz, 1984; Ohlsson, 1984a, 1984b; Snow, 1978, 1981; Snow & Lohman, 1984). As Siegler and Campbell (1990) note:

> The general point is clear: cognitive diagnoses can go seriously awry if they ignore the diversity of people's strategies....That people use diverse strategies is not a mere idiosyncracy of human cognition. Good reasons exist for us to know and use multiple strategies. Strategies differ in their accuracy, in the amount of time needed for execution, in their memory demands, and in the range of problems to which they apply. Strategy choices involve tradeoffs among these properties; people try to choose strategies that enable them to cope with cognitive and situational constraints. The broader the range of strategies that we know, the more we can shape our approaches to the demands of particular circumstances. (p. 4)

Siegler and his colleagues also have created a computer simulation model of strategy choice for children attacking addition, subtraction, and multiplication problems. The model assumes an innate retrieval mechanism plus various backup strategies taught by parents,

teachers, or textbooks. It first attempts to retrieve an answer to a problem based on an associative structure reflecting prior experience with that problem and answer, and applies confidence and search length criteria. When an answer is not retrieved within these criterion limits, a range of backup strategies comes into play. Individual differences in strategy choices also appear, as a function of differences in confidence criteria and in peakedness of associative answer distributions. Not only do empirical data lend support to the simulation model, they distinguish three categories of students for whom different instructional implications follow. "Perfectionists" are highly accurate but take longer than "good students", apparently because of overly high confidence criteria. The "good students" show somewhat lower accuracy but high speed, because they use somewhat lower confidence criteria. It is suggested that both kinds of learners might be helped to tune their performance further--perfectionists by lowering slightly their confidence requirements for stating retrieved answers, since they are likely to be correct anyway, and good students by raising their confidence requirements, thus lessening reliance on retrieval for difficult problems so that back up strategies might come into play. The "not-so-good" students need more substantial intervention to improve accuracy of back-up strategy execution and thus also to raise confidence criteria by sharpening associative peaks for correct answers through further, more successful learning. In other words, Siegler's work supports an aptitude-treatment interaction (ATI) hypothesis. For present purposes, the implication also follows that fine-grain diagnoses of strategy use of this sort, based on simulation modeling, might provide an important adjunct to Case's developmental instructional procedures.

Simulation modeling has become a hallmark of cognitive science. Beyond its use in explicit theory development, it is increasingly turned to advantage in discovering specific

sources of difficulty in school learning tasks that in turn suggest points of improvement for both instruction and assessment. Greeno's (1976, 1978, 1980) work offers several useful demonstrations. In one, Greeno developed a simulation for geometry problem solving that might be considered a model of the objective of successful instruction, i.e., of the performance of an expert (as in Case's step 2) or of a successful student (as in Case's step 3). But the successful model requires three kinds of knowledge: theorems, or propositions for inference; perceptual concepts for pattern recognition, so that appropriate propositions can be retrieved; and strategic knowledge for planning and setting goals in developing a proof or solution. When compared with conventional teaching or textbooks, it becomes clear that the third kind of knowledge in the goal model is typically omitted from instruction. This leads to specific suggestions for redesign of particular lessons, but also to a consideration of whether or not the strategic knowledge component should be taught explicitly, and if so how. The concern is that explicit instruction in all aspects of problem solving may remove the inductive, discovery aspect of learning that is important for some students. This underscores another ATI hypothesis: learners who are able to infer the missing procedural knowledge for themselves do so with special benefit to their own progress, in comparison to the cognitive interference or motivational turn-off that may occur when they are told what to do directly; less able reasoners, however, need explicit guidance in strategic planning because they cannot or do not infer it for themselves (see Mandinach, 1984; Snow, 1982).

In a similar analysis of arithmetic word problems, Greeno and his colleagues (Riley, Greeno, & Heller, 1983) found that successful performance required three types of knowledge structures--one for change of a quantity, another for combination, and still

another for comparison. The ability to discriminate them and relate each to the appropriate sequence of arithmetic operations was essential. But many learners find comparison problems particularly difficult, apparently because this structure is not adequately distinguished, connected to particular operations, or concretized by conventional instruction. This leads to the suggestion that concept discrimination training be made an explicit part of instruction. The training would distinguish the three kinds of problem structures and show how the same arithmetic procedures take on different meanings in the three contexts. Again, however, discovering the appropriate structure for a problem may be an important part of learning for the able student, so explicit instruction might not be best for all; here is another ATI implication. Without question, however, assessment would need to be focussed on the tripartite distinction and the reconceptualization of arithmetic procedures that each implies.

As with the Case and Siegler work, Greeno's studies show the importance of understanding, for particular problem solving situations, the task and goal representations and the bridging strategies that more and less successful students use. Diagnostic assessment of this has to be done in concert with an examination of existing and alternative possible instructional designs.

Learning Assessment and Scripted Tutoring

Our second example is closely related to the first, in the sense that it too concentrates on cognitive structures, strategies, and transfer relations, in families of instructional tasks. But its central focus is a cognitive theory of learning and transfer, and the secondary developmental emphasis comes from Vygotsky's (1978) concept of a "zone of proximal development" rather than from neo-Piagetian stage concepts. The work is that

of Brown and Campione and their colleagues (Brown, Bransford, Ferrara, & Campione, 1983; Campione & Brown, 1987, 1990; Campione, Brown & Ferrara, 1982; Palincsar & Brown, 1984).

Basic concepts and procedures. Brown and Campione first assume that learning abilities are malleable; initial deficits can be remediated, at least for many learners. Learning can be promoted in the short term by carefully orchestrated activities, usually mediated through tutorial dialogue. Performance tasks are designed and sequenced according to a cognitive theory of learning and transfer. Associated with these tasks, a hierarchy of tutorial hints is designed to discover the type and depth of problems that a learner is experiencing as the task sequence and tutoring proceeds. Diagnosis and remediation thus are integrally linked, since diagnostic assessment flows directly from attempts at instruction. The diagnosis then can be used to guide instructional design beyond the range of tasks used in the assessment.

Campione and Brown (1987) apply Vygotsky's theory of proximal development. Learning is mediated by social interaction; an adult assists a child until eventually the latter takes the initiative and assumes responsibility for his or her own learning. Thus, there is a gradual transfer of control over the learning situation from teacher to learner. The zone of proximal development is the difference between the level of performance a learner can attain without instructional assistance and the level that can be reached when guided by a teacher or tutor. The zone provides an initial assessment of a child's learning ability and serves as a guide for subsequent instruction on targeted tasks. Continuous assessment is essential because learning changes the zone of proximal development for each learner, thereby necessitating frequent modifications in both instruction and assessment. Both

within a range of tasks and between ranges, therefore, there are psychometric considerations concerning the measurement of change that need to be explored.

Campione and Brown (1990) first choose a range of tasks in a domain and define a continuum of five levels, from original learning through maintenance to three degrees of transfer. Initially, basic rules are learned with tutorial hints as needed to build learner competence on the first level of problems. Then, at each level the same rules previously learned apply, but the problem context is gradually made more complex by increasing the difficulty of identifying and classifying the problem and of extracting the relevant information from the problem statement. The maintenance level itself offers a transfer practice continuum, in that the identical problem type can be repeated with varying problem statements, in the context of various other problem types. The steps to near and far transfer involve new problem content, the need to produce novel combinations of rules previously learned, and the need to discover additional rules not explicitly taught in earlier phases of the assessment.

The hierarchial structure of tutorial hints is central in the Brown-Campione procedure. It provides a script for the tutor to follow when a performance failure occurs. An example hint structure for arithmetic word problems would be as follows:

1. Give simple negative feedback and prompt to try again.
2. Give working memory refresher, repeating the starting quantity, operation, added quantity, or all.
3. Give numerals as memory aids.
4. Give transfer hint referring to previous problems.
5. Give enumerative strategy hints, suggesting general strategy,

correcting set size, instructing to make a set, facilitating accurate set formation, instructing to count all, and instructing on cardinality.

6. Give complete demonstration and rationale.
7. Give strategic orientation hint, with parallel performance by learner and tutor as model.
8. Abandon ineffective strategy hint.

The tutor moves up the levels of the hierarchy as the learner continues to have difficulties given lower level hints. The number of prompts needed is thus a measure of the degree of help needed. Coupled with the transfer task continuum, a measure of transfer distance also is produced. But qualitative description of learner performance with the hint structure also offers diagnostic information (see Campione & Brown, 1987). The procedure has been used successfully with algebra as well as elementary arithmetic. Some studies have been conducted using series completion and progressive matrices problems of the sorts that appear on tests of fluid intelligence.

Campione and Brown (1990) suggest that the graduated scaffolding of the hint structure and the tutor demonstrations not only build skills in the domain but also aid in the development of more general self-regulatory strategies. The learner models and eventually internalizes the metacognitive as well as the cognitive characteristics of the tutor. The result is a repertoire of multiple strategies, skill in their use, and thus flexible adaptation to instructional opportunities and demands.

An example. Ferrara (see Campione and Brown, 1990) used five levels of tasks in a study of learning assessment in arithmetic word problems for young children. Scripted

tutoring with a hierarchial hint structure followed the procedure previously noted. A knowledge test covering number-word sequences in counting, appreciation of the contexts in which counting occurs, and the actual performance of counting, was used as a pre- and post-measure to assess gain.

There were several important results. Scores based on the hint structure and transfer distance were better predictors of learning gain than were ability and knowledge measures, and accounted for unique variance in these predictions. Furthermore, substantial variability was noted among learners: a significant number were seen to develop strategies for solving the word problems that differed from those taught in the structured tutoring; they also showed more initial knowledge but also more gain. IQ scores did not predict gain in this group, but did predict for the remainder of the sample. Thus the role of initial ability and knowledge in strategy acquisition was seen to vary with the development of skill in the domain. If it is the case that skill acquisition depends more on general ability among those with less initial knowledge, then these results correspond to other theory and research suggesting that general ability differences influence mainly the early, declarative knowledge phases of skill acquisition (Ackerman, 1989).

The Campione-Brown work suggests that their learning and transfer measures can add significantly to assessments provided by conventionally designed ability and knowledge tests. The learning measure is thought of as "learning efficiency"--the inverse of the explicitness of help a person needs in order to learn. The transfer measure is "transfer propensity"--the extent to which what is learned can be applied in increasingly novel situations, across a range of tasks reflecting transfer distance. The latter appears to be the most sensitive measure of individual differences and the most general measure of learning

outcome to be derived from this work; it distinguishes among students on flexibility in use of knowledge even after all have learned the original arithmetic rules to the same criterion.

These results also can be interpreted to support the view that measures of flexible processing in transfer performance are needed in assessment, along with measures of growth in domain-specific knowledge. This sort of processing is seen as involving the operation of metacognitive, self-regulatory functions. Monitoring of progress in learning thus needs to include regular checks on flexible use of the skills being acquired, to identify strengths and weaknesses in these functions. These differences between more and less efficient learners can be seen in think-aloud protocols Campione and Brown collect. More able learners spend time planning and analyzing, checking their reasoning, and monitoring--and they have repair strategies to use when things go wrong. Less able learners, in contrast, try out rules randomly, moving quickly to solutions, without analysis or reference to previous problems. The instructional implication is that "training for transfer" that focuses explicitly on these functions should be included in teaching or tutoring when they are diagnosed to be deficient.

Related research. Much of the cognitive and developmental analyses cited under the previous example applies here as well. In addition, there are a variety of related approaches arising from clinical research with retarded and disadvantaged children (see Lidz, 1987). Aside from Brown and Campione, the most comprehensive programs are probably those of Budoff (1987a, 1987b) and Feuerstein (see Feuerstein, 1979; Feuerstein, Jensen, Hoffman, & Rand, 1985; Feuerstein, Rand, Jensen, Kaniel, & Tzuriel, 1987).

Budoff (1987a) seeks to assess learning potential using a test-train-retest sequence to obtain gain scores. The tests, for the most part, are standardized fluid intelligence

measures. The training focuses on test problem-relevant basic skills and strategies, although experiments have also been conducted with elementary mathematics and science materials. Results seem to indicate that the procedure can identify learners in groups classed as low IQ who are able to learn in other educational settings. Various interactions with other personal characteristics have also been reported, however, and the training also shows marked effects on some personality correlates of learning. Current work aims at designing methodologies to translate learning potential measures into individualized prescriptions for treatment. Budoff believes that his standardization of testing and treatment is a key advantage, because individualized clinical interactions aimed at high-order cognitive skills cannot readily be evaluated for generalization and transfer to conventional educational and other everyday life settings without such standardization.

The Feuerstein approach involves just such intensive, clinical interactions with the learner in individualized teaching sessions. It is based on the view that mediational deprivation during development reduces modifiability of performance (Feuerstein et al., 1985). Cognitive processes develop through normal everyday interaction with the environment but also through explicit exposure to mediated learning experience. The latter occurs when an adult mediates between a child and a learning task, by modifying the stimulus in some way for the learner. Through this mediation, learners increase their capacity to benefit from exposure. Inadequate cognitive development and decreased modifiability occur when there is insufficient mediation. That is, even after repeated exposure, deprived learners appear to be unable to adapt to new situations if appropriate mediation is not provided. However, through controlled instruction, known as Instrumental Enrichment, cognitive modifiability can be increased.

According to Feuerstein et al. (1987), Instrumental Enrichment is a flexible, individualized, and interactive program focusing on fluid rather than crystallized abilities that are accessible to change. It includes several series of exercises and systematically ordered didactic techniques. The intent of the program is to increase a learner's capacity for modification by providing mediated learning experiences to correct deficient cognitive functions. The increased capacity for modification enables the student to adapt more readily to varying situations. It targets cognitive skills for particular tasks, and seeks to promote intrinsic motivation, improved working habits, insights into cause-effect relations, and self-concept as a learner. The program consists of 20 instruments, each of which targets one or more cognitive functions. A cognitive "map" depicts the cognitive functions presumed to be involved in each task performance. The map is a system for classifying cognitive activities according to content, modality, phase of cognitive functioning, cognitive operations involved, complexity, abstraction and efficiency. Instructional units then focus on a limited number of these cognitive functions, but also attempt to reinforce other functions established in previous tasks. With initial changes in cognitive functioning, new situations can be formulated from the map to bring the learner toward more adaptive performance.

The field represented by all this work has come to be called "dynamic assessment". This term refers to a collection of procedures designed to assess and promote learning ability more or less simultaneously, usually in the context of an individualized intervention to remediate learning difficulties or low intelligence test performance. Indeed, early work in this direction grew as much from the need for improved clinical treatment of learning ability problems in special populations as it did from dissatisfaction with diagnoses based

on conventional "static" tests. Unfortunately the dynamic-static contrast has become a rallying cry that obscures important differences in methodology among its proponents, as well as important research issues for psychological and psychometric theories of learning and intelligence more generally. In a later section, we reject the "dynamic-static" terminology. Without detailing our reasons here, we prefer to refer to this line of work more directly as "learning assessment". But the confluence of theory and research on several of the procedures has now progressed sufficiently to show that the basic concepts merit sustained attention, whatever the label used.

Computerized Tutoring as Assessment

Our third example comes from the development of computer-based tutors that use micro-level assessments to guide instruction. A number of research projects aimed at this objective now exist (see e.g., Anderson, 1988; Gitomer, 1988; Gitomer & Van Slyke, 1987/88; Lesgold, Lajoie, Bunzo, & Eggan, in press; Mandl & Lesgold, 1988; McArthur & Stasz, 1989; Polson & Richardson, 1988; Psotka, Massey, & Mutter, 1988; Shute, Glaser, & Rughaven, 1989; Wenger, 1987).

Basic concepts and procedures. Burns and Capps (1988) suggest three criteria that distinguish intelligent tutoring systems from previous forms of technology-based instructional systems. The first is that the system must itself contain the complete content domain in order to make inferences about a student's knowledge with respect to it. Second, the system must be able to detect different levels of that knowledge. Finally, the instruction must specify strategies aimed to decrease the differential between expert and student performance. For detailed discussions of these facets of tutor development, see especially Wenger (1987) as well as Polson and Richardson (1988) and Winne and Jones

(in press). We briefly describe here some of the key concepts.

A first critical feature is the expert module that contains the content knowledge. Both the concepts and the procedures of a domain and the interconnectedness among concepts and procedures must be represented within the system in fine detail. Modeling expert knowledge in a domain is an enormously complicated problem, and major efforts in artificial intelligence research have been invested in just this part of the task.

The second key feature is the student module, the knowledge states and structures that represent a student's understanding of a domain. Van Lehn (1988) distinguishes student declarative and procedural knowledge, arguing that the easiest form of knowledge to represent is flat procedural knowledge, followed by hierarchial procedural knowledge. Declarative knowledge is the most difficult modeling problem, in part because it should include all relevant background knowledge, or "common sense", as well as explicit domain knowledge. Furthermore, as Anderson (1988) notes, the system needs not only theories of performance but also theories of learning to understand the cognitive processes operating within a domain that lead to different knowledge states.

A third component is the instructional module that chooses and implements the teaching steps necessary to move the student toward the goal. Halff (1988) notes that the instructional design chosen for a tutor depends heavily on what assumptions about the nature of learning, teaching, and subject-matter are adopted. A representation of the content must be chosen. This determines the sequences of concepts to be represented, and all subsequent steps including assessment steps in the tutorial interaction. In short, a theory of instruction and assessment in the domain is also needed.

To date, most intelligent tutor developments seem to fall along a continuum and

have been based on one of three theories of learning and instruction. One embodies the principles of inductive or discovery learning. The computer provides a learning environment to be explored. It might simulate natural phenomena, for example, and provide instruction in tools to use in studying the simulation. But the learner is free to follow idiosyncratic strategies in discovering the goal concepts. The second type follows the rules of direct, didactic instruction. The computer directs the instructional steps, provides immediate feedback and correction, and constrains learner attention to avoid digressions. Programs by Collins and Brown (1988), White and Frederiksen (1987), and White and Horwitz (1987) exemplify the first approach. Anderson's work exemplifies the second (see Anderson, 1988; Anderson, Boyle, & Reiser, 1985; Anderson, Boyle, & Yost, 1985; Anderson, Conrad, & Corbett, 1989). Of course, mixtures of principles from each tradition are possible (see Frederiksen & White, 1990); there might be free exploration coupled with critical advisory feedback. An example of such a system is Writer's Workbench (Frase & Diehl, 1986; Macdonald, Frase, Gingrich, & Keenan, 1982).

Other aspects of tutor design concern assumptions about the instructional environment in which the system is to be used (Burton, 1988) and the human-computer interface that will be most appropriate (Miller, 1988). Critical implementation (Johnson, 1988) and evaluation (Littman & Soloway, 1988) issues must also be confronted.

An example. The system we choose to illustrate one mixture of principles for intelligent tutoring is SHERLOCK, developed by Lesgold and colleagues (Lajoie & Lesgold, 1989; Lesgold, 1988; Lesgold et al., in press). Its purpose is to improve the troubleshooting abilities of Air Force electronics technicians beyond their regular training. SHERLOCK is considered a coached practice environment rather than an intelligent

tutoring system in the strictest sense, but it exemplifies several key features for our purposes here. It is especially valuable for its approach to diagnostic assessment.

The design of SHERLOCK is based on a system of hints rather than direct instruction. The hints are structured to provide varying degrees of specificity depending on variation between expected and actual levels of the individual's performance, and can be given directly or at the student's request. SHERLOCK first generates a general hint, thus stimulating thinking rather than providing answers. In this way, it is more similar to models of human tutoring (see e.g., Swanson, 1989) than are most other computerized tutoring systems. One purpose of the hint system is to promote collegiality by allowing for collective group discussion (since all trainees can receive the same problems regardless of their ability level). This is an important feature, especially in view of demonstrations that social interaction helps develop higher levels of problem solving abilities (see, e.g., Schoenfeld, 1985). Also, however, the hint structure can provide a highly individualized diagnostic assessment similar to that used by Brown and Campione. This is a critical difference between SHERLOCK and systems in which instruction is highly directed.

SHERLOCK also incorporates knowledge and skills from outside as well as inside its immediate domain, thus providing a context-situated learning experience (see also Brown, Collins, & Duguid, 1989b). Until recently, relatively little emphasis has been placed on this aspect of the learning environment. But several researchers (see Greeno, 1988; Rogoff & Lave, 1984) now believe that the problem of transfer of training is rooted in the lack of attention to this issue. By merging in-shop and other related problems in simulations, SHERLOCK provides cross-training among several job specialties and thus attempts to transfer skills and knowledge to the real world job.

Still another critical design feature of SHERLOCK is its multiple viewpoints on learning. It is not dependent on one particular choice of student model as are many tutoring systems (Lajoie & Lesgold, 1989), but rather provides an adaptive environment that takes into account two types of student models--performance and competence. Dynamic modeling of both competence and performance allows SHERLOCK to determine the amount and level of help needed by each student. The competence model reflects achievement of the curricular goals of SHERLOCK. The performance model reflects the student's problem solving activity within the problem space; it reflects the level of assistance needed using a performance-prediction differential.

The heart of the performance model is the effective problem space, which consists of a network of nodes reflecting partial solutions and links that represent transitions between solution states. Although separate problem spaces are generated for each problem, simulations compare several partial solutions to expert performance. This is part of the basis for generating hints. The model is a recording-keeping device that monitors the student's progress through the system. It also provides a summary of the hypotheses and performance generated by the student.

An important aspect of the performance model is its adaptation of next instructional steps based on student aptitude information from outside as well as inside the immediate performance domain. Lesgold (1988; see also Lesgold, Bonar, & Ivill, 1987) describes how this kind of diagnostic decision-making works. In deciding on the next step, the program uses information on the present state of the student model, in relation to the curriculum goal structure, and a set of instructional design rules. One such rule might be: Given this student's learning state, work next on voltage measurement skills associated with Ohm's

Law that are in a probable (as opposed to strong or absent) state. But the program also chooses the next step using aptitude information--a second rule might be: Use diagrammed circuits with simple computations because this student is strong in figural comprehension but weak in numerical skills. The next instructional step thus is chosen or designed using a list of such constraints; in principle, the constraint list could include information from any internal or external source.

The competence model embodies Lesgold's (1988; Glaser et al., 1987) conception of a goal lattice layer for curriculum design.

> "The goal lattice layer is a lattice structure in which are encoded a number of goal hierarchies, each corresponding to a fundamental viewpoint of the task of teaching the course content....This multiple viewpoints approach, incidently, has implications for what constitutes an appropriate course, in terms of the completeness, coherence, and consistency of its curriculum lattice....The course is coherent, in that each simple lesson is relevant to all of the viewpoints we have taken. It is locally complete, in that each viewpoint seems to be completely teachable with the set of simple lessons....It is globally complete to the extent that the viewpoints represented include all of the viewpoints routinely held by experts and any others that are important to learning the domain content. Finally, it is relatively consistent, in that the prerequisite relationships all run in the same direction. There are no cases where lesson X is prerequisite to lesson Y from one point of view while lesson Y is prerequisite to lesson X from another". (Lesgold, 1988, p. 126, emphasis in original).

A goal lattice, then, is a network of lessons--a kind of curriculum map, with routes connecting key points of interest. Each node is a lesson aimed at one or more component

concepts or skills to be acquired. Combined with the performance model, each node is also represented by one or more performance tasks. Scores on these tasks can be used to guide instruction, but also to represent the state of each learner's progress regardless of which teacher or expert viewpoint, or presumably which of several possible routes through the course is taken. The properties of coherence, completeness, and consistency as described by Lesgold appear to be important criteria for the design of reference tasks in such a map. Each task should be relevant from all teacher or expert viewpoints. The collection, or map, should be complete in that no important subgoal from any viewpoint is omitted. And the map should be consistent in that the tasks do not take opposite prerequisite orders from different viewpoints.

Related research. Shute (1990) conducted a large scale study of SHERLOCK to demonstrate the importance of adapting further intelligent tutor development to assessments of student differences. She manipulated feedback conditions experimentally to contrast complete explication for each problem (deductive feedback) and feedback that left relationships to be inferred (induced feedback). She also included measures of cognitive abilities before instruction as well as exploratory behavior during instruction. The results showed interactions with both student variables, suggesting that deductive feedback was superior for more able learners in this sample who exhibited low exploratory activity during learning, but not for other students. Inductive feedback served no one well. Although her data analysis is not yet complete (at the time of this writing), Shute's research provides a valuable implication for our purposes: tutorial instruction of this sort needs to be geared to assess and adapt to student aptitude differences exhibited before instruction and to learning activity differences exhibited during instruction. Both of these sources of

information lie outside of the competence models usually used in tutor design.

A small-scale correlational study by Snow, Wescourt and Collins (1979; see also Snow, 1980) yielded the same implication in relation to instruction in BASIC programming. Both cognitive and personality aptitudes correlated with learning rate and posttest measures. But so also did various indices of learning activities during progress through the program. The BASIC Instruction Program (BIP) was one of the first adaptive computer systems invented (see Barr, Beard, & Atkinson, 1976). The learning goals and the student model is represented as an interconnected network of skills and knowledge. BIP selects the next problem according to the student's position in the curriculum and continuously updates the student model to reflect past performance in the program content domain. Many other current programs use a variation on this theme. But the system also generates a number of indices that reflect learning strategy and activity differences, time allocation variables, requests for help, and various other learning process characteristics that lie outside of the content domain (see Kyllonen & Shute, 1989). It does not appear that this is an important source of information for adaptive instructional purposes.

Several other intelligent tutoring systems offer additional design ideas worthy of note. We describe only a few here.

Some systems attempt detailed diagnosis of student errors by incorporating bug libraries and production rule modeling of student performances. Examples are BUGGY (Brown & Burton, 1978), DEBUGGY (Burton, 1982) for arithmetic and PIXIE (Sleeman, 1987) for algebra. In DEBUGGY, procedural knowledge is modeled as a production system. Some errors students make on computation tasks are then seen to be predictable consequences of incorrect, or "buggy" production rules. Diagnosis of student bugs is

achieved by comparing the procedural errors made by a student over a series of problems with the system's bug library.

PIXIE attempts to escape the limits of a fixed library of bugs by detecting and adding new ones. Student performance is represented as an ordered set of production rules. Variants from correct rules are referred to as mal-rules. Although the conditions under which these mal-rules apply correspond to those of the correct rules, the actions dictated are incorrect. Unlike the systems that rely on bug catalogs, PIXIE will hypothesize a new mal-rule to represent the student's action if it cannot otherwise account for it. Sleeman makes a further distinction between manipulative and parsing mal-rules. Manipulative mal-rules occur when the student fails to apply a rule correctly even though knowing the rule. Parsing mal-rules are more serious errors reflecting a significant lack of understanding of the rule. This sort of distinction appears to have additional value for diagnostic assessment.

WUSOR, a computer coach for a reasoning game called WUMPUS, bases its student model on a check list of skills to be noted if exhibited. Goldstein and Carr (1977; Carr, 1977; Goldstein, 1979) developed a four-part overlay model using a production-rule representation of knowledge states designed to indicate the presence or absence of specific skills. The student's understanding at each state is represented as exhibiting some or all of the components of how an expert performs in each specific situation. It therefore exemplifies a computer version of part of the Case-Bereiter developmental approach.

A contrasting system is Smithtown (Shute et al., 1989), a discovery environment for microeconomics. The goal of Smithtown is for the student to induce knowledge of microeconomics and general problem solving skills from observation, discovery, and

practice through experiment. Unlike the more directive systems, Smithtown does not impose a particular curriculum on the learner. Instead, the student supplies problems, questions, and hypotheses (Kyllonen & Shute, 1989). Smithtown then monitors the student's actions, maintains a history, and provides coaching on buggy behavior and errors of omissions. Discovery learning in microworlds of this sort may be uniquely useful for diagnostic assessment of understanding beyond explicit curriculum goals.

Other ongoing tutor development efforts include algebra (Anderson, 1988; McArthur & Stasz, 1989), Pascal programming (Johnson & Soloway, 1984; Shute, 1989), electronic troubleshooting (Frederiksen, White, Collins, & Eggan, 1988), geometry (Anderson et al., 1985). There is also a system for scoring Pascal programs generated in other instructional settings (Bennett, Gong, Kershaw, Rock, Soloway, & Macaladad, 1990; Braun, Bennett, Frye, & Soloway, 1990; Braun, Bennett, Soloway, & Frye, 1989). However, only recently have researchers turned their attention to evaluating the impact of such systems on cognitive performance more generally (e.g., Lesgold et al., in press; Shute, 1989; Shute et al., 1989). Such evaluations are an important next step in intelligent tutor development and implementation.

Learning Progress Assessment in a Curriculum

Our fourth example is the development of learning progress systems as conceptualized by Bunderson and his colleagues (see Bunderson et al., 1989; Forehand & Bunderson, 1987a, 1987b; Murphy & Bunderson, 1988). The aim here is to provide assessments of learning progress or student trajectory over whole courses of instruction. The assessment tasks are themselves also learning experiences placed at key points in a curriculum domain.

Basic concepts and procedures. A Learning Progress System (LPS) is not conceived to be a curriculum or an instructional program in and of itself, but it is intended to span a whole course as conventionally defined. At base, an LPS is a network of performance tasks designed to reflect the goals of an instructional program, and the intermediate objectives or milestones recognized as indicating learning progress toward these goals. The individual tasks in the network, and scores derived from them, are designed to refer not only to immediate instructional objectives but also to alternative next steps for instruction toward further goals, given the state of learner progress they indicate. Such tasks are thus called reference tasks. The network of reference tasks for an instructional domain forms a learning progress map that can be used by teachers, learners, and parents to chart progress and to communicate about instructional steps and goals. The tasks in the map form continuous scales in the sense that they are calibrated for difficulty and placed in the instructional sequence to permit smooth steps from one to another, rather than to require large disjoint leaps. The reference tasks, their scoring, and their calibration values, are designed to show the component knowledge and skills required for learning success, the reasons for learning difficulties, and the relevance of such performance to real-world situations. In short, the system is intended to be valid for instructional diagnosis and adaptation, but also for communication both within and beyond the school.

A first step for any LPS is to reach a carefully articulated set of instructional goals in some domain and at least an outline of intended content for an instructional program to reach the stated goals.

The next step is construction of the reference tasks. This starts with a table for each goal or closely related set of goals, wherein suggested tasks can be exemplified and

initial hypotheses about their psychological properties can be listed. The examples are categorized roughly as to presumed difficulty level. The tasks should exhibit clearly their relevance to the goals to which they are attached. They should refer to or simulate actual activities expected later in academic work, in college, in employment, or in everyday life. They should also require multiple, constructed responses that can be scored to assess each learner's particular instructional needs and progress. Specialized instructional or coaching materials should be identifiable that address each component of skill or knowledge required by each task. The tasks should be reusable with practice, or include parallel materials to allow continued practice. Successful performance should be observable by the learner--that is, another person should be able to model clearly successful performance for the learner to watch. Taken together, the tasks should span a wide range of presumed difficulty and provide series of milestones through the course.

A wide range of stimulus material then is assembled or drafted to fit the table of specifications. Some principles of conventional test development are used. But candidate reference tasks also require review by teachers and tryout by developers in small-scale experiments to analyze prerequisite knowledge and skills, likely errors or misconceptions, and the nature of learning progress exhibited by the task performance. The proposed scoring system for the task is geared to reflect these key features of the performance.

The table of tasks is structured to form a map that reflects the instructional goals and intermediate objectives in the domain. The map indicates a learner's current position and alternative routes to a further goal, and should help teachers in selecting the most appropriate next instruction for an individual learner. The construction of this map also involves assumptions about learning progress and outcomes that need to be verified by

review and experiment. In particular, empirical studies are needed to establish the calibration of the reference tasks statistically. Difficulty scale values thus derived must be examined across tasks to compare the apparent requirements of tasks at different levels of performance.

Evaluation studies also are conducted on portions of the network of tasks or on the map as a whole. These examine the placement of tasks with respect to learner readiness for instruction along different routes toward a goal. They also check and improve the effectiveness of coaching associated with particular tasks or transitions between tasks. Ultimately, these studies verify the effectiveness of series of reference tasks and associated instruction in providing alternative routes toward goals for learners showing different profiles of current strengths and weaknesses.

An example. Although no complete LPS yet exists, one concrete example that approaches the LPS scheme is under development in the area of literacy among young adults. It is based on tasks used by the National Assessment of Educational Progress (NAEP) in a national survey. NAEP used lists of performance tasks to form separate prose, document, and quantitative literacy scales. Difficulty values for each task designated the point on that scale at which individuals with that level of proficiency have an 80% probability of responding correctly. For example, relatively difficult tasks on the prose, document, and quantitative scales might require, respectively: identifying particular information in a lengthy news article; using a bus schedule to select the appropriate bus for given departures and arrivals; and determining the amount of interest charges from a loan advertisement. Relatively easy tasks on the three scales, respectively, might require: writing a letter stating that an error has been made in billing; entering personal information on a

job application; and entering and calculating checkbook balances (for details on actual tasks and analyses, see Kirsch & Jungeblut, 1986). These tasks and scales seemed to function well for the purposes of NAEP. To make them into reference tasks for LPS, however, several further steps are required.

First, the scales gloss over subsidiary distinctions that may be important for learning diagnosis and instructional planning. For example, the prose scale contains three separate subcategories of reading comprehension skill: locating information in a text, reproducing and interpreting text information, and generating a theme or organizing principle from text information. Also, the tasks differ in their demands on reading, writing, and oral production. Cognitive analysis of the tasks to identify key knowledge and skill components would undoubtedly identify other distinctions important for scoring and for instructional purposes. Kirsch (1987; see also Kirsch & Mosenthal, 1988; Kirsch, Mosenthal, & Rock, 1988) has begun research in this direction. For each task, variables are identified that might influence the cognitive complexity of performance, e.g., the number of features in the question that must be matched to the document, the degree to which wording of question and document correspond, and the number of plausible correct answers. These are then used to examine sources of difficulty for the NAEP sample and for targeted subgroups.

Second, the scoring system for these tasks must be elaborated. At present, the plan for LPS is to use three-category scoring: competence demonstrated, needs instruction, or not ready for this task. A learner identified as needing instruction can then be given coaching addressed to particular component skills or misconceptions. The scoring algorithms that would make such classifications have not yet been developed. Moreover, it is likely that a more detailed diagnosis will be needed if the LPS is to guide teacher

choice among instructional alternatives. Careful study of the mass of response data collected by NAEP on these tasks, however, provides a start toward designing the needed algorithms.

Third, of course, the instructional and coaching materials, and the parallel forms to allow practice on each task, are not yet in hand. A substantial investment in production and evaluation of these materials is thus another important next step.

Related research. Young adult literacy is one important domain chosen for LPS development. Since much is gained by comparing developmental experience across domains, the program is also engaged in collecting and designing reference tasks for middle school science (Gong, 1988). Beyond this, there are dozens of school learning tasks, and also ability and achievement tests, that have been subjected to cognitive analyses of the sort needed to transform them into reference tasks for potential LPS. Indeed, the tasks studied so far in Case's developmental assessments, Brown and Campione's learning assessments, Lesgold's SHERLOCK, and in many other related developments, are candidate reference tasks for LPS. So far, however, no further research has sought to string them together into the sorts of networks or maps envisioned here.

One related project has produced a prototype system for college-level remedial instruction in reading, writing, and study skills (Forehand & Rice, 1988). It consists of computerized exercises that call for constructed response, provide feedback and practice, and branch to instruction focussed on particular diagnosed weaknesses. The program is also designed for interactive use by students and teachers. It would appear that this system could readily be combined with the adult literacy tasks studied by Kirsch (1987) and his colleagues to produce a prototype LPS.

Basic Issues for Research and Development

We chose to review these four systems and some related research in order to use their similarities and differences to illustrate basic issues needing further attention. The aim is to reach a general research and development agenda, of use no matter which of these or various other approaches one might choose to pursue. We thus do not make itemized evaluative comparisons among the approaches reviewed. The four are still in early stages of development, and they are not competitors. Indeed, they could be further developed to coexist, even to complement one another, in the same instructional program. We consider this possibility at several points below because we see their joint development and use as a highly valuable step.

In the present section, we summarize the issues raised by our examination of the example programs. In later sections, each issue is addressed in more detail. To conserve space, however, we limit cross-referencing and recounting of previous discussion. Also, we use acronyms when referring to systems for integrating instruction and assessment in general (SIIA), and when referring to our example approaches--the Case-Bereiter Developmental Assessment System (DAS), the Brown-Campione Learning Assessment System (LAS), Lesgold's Computerized Assessment System (CAS), and Bunderson's Learning Progress System (LPS).

The issues can be grouped into six sets and summarized in question form, as shown in Table 1. The resulting list is an elaboration of the four questions with which we began.

Insert Table 1 about here

First, the example systems take different stances with respect to defining instructional goals, domains, and treatments, and one could take still other stances. There are questions about whether the systems require or assume hierarchically organized goals, whether they accommodate different expert viewpoints regarding goals, and whether they fit domains other than the well-structured character of mathematics and science. The systems differ in their explication of accompanying instructional treatments and also aim at different scale or grain-size levels.

A second set of issues concerns the design and use of assessment tasks, tests, and other indicators of learning. Some systems choose tasks that are intrinsic to instruction, that embody end-goals of instruction themselves. Some design adaptive instructional and assessment tasks. Some could accommodate a variety of tasks found naturally in classroom instruction.

Third are questions pertaining to teacher roles. The systems differ in the degree to which they fit well with conventional classroom practices and in their provisions for using teacher observations and judgments, for teacher training, and for promoting teacher-student and teacher-parent communications.

Fourth, the systems differ in theoretical perspectives on learning, development, and individual differences. They reflect somewhat different concerns about the use of errors, the promotion of transfer, and the nature of expertise.

Fifth is the definition of diagnosis reflected in each system. Diagnosis arises from different sources and serves somewhat different purposes in each system.

Finally, system evaluation and implementation involves somewhat different issues in each case, since the systems are presently at rather different stages of development.

The sets of issues overlap and interact. In each case, there is an intermingling of theoretical research issues and more technical issues of engineering design and development.

Instructional Goals, Domains, and Treatments

Traditional definitions and boundaries. Educational psychologists usually have taken instructional goals as given by curriculum specialists and teachers. These goals usually are cast within the traditional curriculum structure of school and college education. Past attempts to improve the school curriculum also have been cast in this framework. So, for example, the College Board publishes a booklet designed to upgrade student preparation by showing students, parents, teachers, and policy-makers what the new consensus goals of high school education should be for college bound students. It lists these goals in various areas of academic competence and domains. It also implies a kind of mastery gradient, at least in mathematics, based on breadth rather than depth considerations, For example: all college bound youth should possess "knowledge of relations, functions, and inverses" and "the ability to graph linear quadratic functions and use them in the interpretation and solution of problems"; but college entrants expecting to major in science, mathematics, or engineering need more extensive proficiency, i.e., "knowledge of ... polynomial, exponential, logarithmic, and circular functions" and "the ability to graph ... and ... use them...." (College Board, 1983, pp. 21-22).

Other pronouncements aimed at reforming primary and secondary education for all students, not just for the college bound, are couched in broader language. But they also stay within traditional domains and, for the most part, use some combination of breadth of coverage and what has come to be called "minimum competency" to characterize

goals (see e.g., Boyer, 1983; National Commission on Excellent in Education, 1983; U. S. Department of Education, 1984). As useful as these efforts may be for some purposes, they tend to slight other possible definitions of instructional goals and domains--definitions that would allow the design of assessment and perhaps also instruction to focus on both broader and deeper goals.

For example, SIIA might be particularly useful when targeted at the interstices between traditional disciplines, or across a range of domains within disciplines. Present curriculum designs and thus current assessment instruments are particularly weak in this regard. Examples of such interstitial areas might be the uses of mathematics in science, computers, auto mechanics, or family economics, or the interrelation of history, art, and geography, or the interdependence of reading and writing across many domains. Some of the more innovative university programs now have this character. Examples from Stanford University are programs in: values, science, and technology; human biology; international conflict; and symbolic systems. Some high schools have occasionally tried such programs. Interstitial SIIA might go a long way toward showing the relevance of one domain to another, thus promoting motivation as well as transfer, while still producing competence in each domain.

It is particularly the deeper goals of education that traditional instruction and assessment do not adequately address. Rather than defining higher levels of proficiency as increased breadth of knowledge and ability across the facets of a discipline, or as minimum competency in each, new research on SIIA might better serve educational improvement by focusing on increased depth of knowledge and ability within one or a few disciplinary facets. For example, in precollege mathematics, new definitions of curriculum

goals emphasize teaching for understanding, real-world applications, investigation, reasoning, and modeling (National Council of Teachers of Mathematics, 1989). Instructional treatments are also to be reformulated. Constructive learning should be emphasized. Classroom activities should be more varied, encouraging student initiative and hypothesis-testing, employing a variety of pedagogical formats with heavy emphasis on computer and calculators. Students should assume active roles also through project work and group assignments. The report has been endorsed by many major educational and scientific organizations (see also American Association for the Advancement of Science, 1989; Murname & Raizen, 1988).

Recent advances in instructional psychology also move in this direction. For example, Greeno's (1986b) review of research on mathematical cognition suggests that the level of mastery typically sought by instruction in mathematics has been something like: "Know the concepts and techniques of mathematics and be able to apply them to problems where appropriate". This is essentially the language of the College Board and many previous reports. But this level seems to reach only an "advanced beginner" level in progress toward expertise. As Greeno shows, cognitive analysis and simulation studies in a variety of elementary and secondary school mathematics tasks have uncovered important structural and process features of such levels of knowledge and suggested instructional improvements as a result. But the research also prompts questions about the adequacy of that goal, for both instruction and research. It is seen that students can achieve this level of knowledge without acquiring deep understanding of mathematical principles or becoming able to reason productively using them in other situations; their knowledge is not generative or transferable.

New psychological goals of instruction. The new work thus adopts a deeper,or higher, psychological goal: "... in effect, it proposes that students should learn to understand and reason in mathematics as mathematicians understand and reason" (Greeno, 1986b, p. 9). They should learn the significant components of mathematical practice, not just the language, facts, concepts and techniques. These include (according to Kitcher, 1984; see Greeno, 1986b): the kinds of questions that are understood as meaningful or useful, the methods of reasoning that can support conclusions, and the meta-level viewpoints that characterize the goals and structures of mathematical knowledge. Such an instructional goal reflects a level of expertise well beyond present practice in mathematics education and also most past research in mathematics cognition. By choosing such a goal, SIIA development might advance both theory and practice substantially, and build a rather different kind of bridge between high school and college, or high school and work, than that promoted in past educational reforms. At the least, the implication is that SIIA should not be cast on outmoded conceptions of curriculum goals and instructional treatment. It does seem that much past research and practice in assessment has made this mistake.

To adopt the higher goal, however, does raise new complications. It seems to mean choosing a few facets of mathematics, or whatever discipline, and concentrating on the full range of expertise in just these. Greeno (1986b) and the mathematicians he quotes do not seem to say so, but it may be both theoretically and practically impossible to build instruction or assessment that produces this level of knowledge and ability in all domains. Palincsar (1989) made this point in commenting on apprenticeship as the instructional treatment of choice in developing expertise. The College Board, for example, lists computing, statistics, algebra, geometry, and functions as the key facets of mathematics, and

this is only one of six disciplines. Presumably, if one chooses to concentrate on mastery in a single facet, one assumes that a student who becomes a master of that one trade can more readily become at least a jack of the others, at least within a discipline. We do not yet have evidence that transfer of this sort happens.

Furthermore, each domain or discipline has its own structure. Each has a central conceptual constitution around which its knowledge base of facts, concepts, and principles is organized. Each also has its rules and techniques for questioning, reasoning, and judging the value of statements in its phenomenal domain. Schwab (1962, 1978) called the first a discipline's substantive structure, the second its syntactic structure. The essential goal of instruction in each discipline, for Schwab, is to communicate these central structures and to bring students to understanding and appreciation of them. Schwab makes the same point for all disciplines as Kitcher and Greeno make for mathematics. But he also shows how concepts apparently held in common by different disciplines carry subtle differences in meaning within each structure; fundamental misunderstandings and confusions arise when instruction does not explicate these differences clearly. This comes back again to the goal choices noted above. Unfortunately, much conventional instruction ignores these deep structures, subtleties and interstices, concentrating instead on the lists of content terms, facts, and methods that "cover" the domain. The problem for the student is thus made immensely more difficult (Shulman & Ringstaff, 1989). So is the problem for assessment design.

Crossing these considerations with current understanding in cognitive psychology, it nonetheless follows that SIIA development needs to identify and focus upon the central conceptual structures in each domain that are essential for deep student understanding and

the conceptual similarities and distinctions between domains that are likely to cause confusion or promote transfer. Of importance also is the identification of key points of juncture and transition in the organization of a domain--the joints at which experts divide and chunk their field, the domain concepts that may be inherently difficult or counterintuitive, and the relations between domain concepts and everyday concepts that may yield malfunctional analogies or misunderstandings. Further, close attention needs to be given to the ways the domain is typically taught and the instructional goals that are typically sought; these may produce additional difficulties which, though not inherent, will be prevailing in the prior knowledge of students and thus hard to remove.

Targets of difficulty. All these considerations suggest that there are targets of difficulty--concepts or skills that are hard to teach and learn, but that are considered centrally important to student progress. Such difficulties might be inherent in the content domain or instructional medium or they might be accidental. Some progress in identifying and understanding how these targets of difficulty occur has already been made by work at Harvard's Educational Technology Center (ETC, 1988). ETC began with four assumptions: (a) there are concepts and knowledge central to understanding a discipline; (b) intensive examination of a discipline will yield information about useful teaching approaches; (c) technology should be used selectively to facilitate teaching and learning of specific content; and (d) technology-based materials should be integrated gradually and rationally into existing course materials and procedures. ETC research then examined teaching, learning, curriculum, and implementation problems, including students' and teachers' perceptions about content, in precollege science, mathematics, and computer instruction. Using these analyses, ETC developed "metacourses" as alternatives to traditional instruction focussed

on the identified targets of difficulty, within and between domains (e.g., Perkins & Martin, 1985; Perkins, Martin, & Furady, 1986).

The development of SIIA ought to take advantage of this approach, and of the prior experience of ETC in identifying and analyzing targets of difficulty. In any particular domain it ought also to take advantage of whatever other forms of cognitive task analysis and simulation modeling have been conducted in that domain. In mathematics, there are dozens of well executed studies of tasks central to the school curriculum. In physics and in reading there are fewer, but still many. In other domains, there are at least isolated examples. Work on SIIA should start by identifying these tasks and arranging them into some form of network in relation to the curriculum structure in that domain. These would then become important starting points toward defining instructional goals and relations among goals in psychological, not merely logical terms. But it should pay particular attention to arrays of such tasks that seem to reflect the deep conceptual and procedural changes that lead toward expertise.

None of our example developments seems to have begun this way, at least not explicitly. This is understandable because the networks of central conceptual structures, key joints, and inherent or apparent difficulties within and between school instructional domains have never been constructed, at least not in psychological terms. Nor have the instructional tasks so far studied psychologically, in mathematics for example, been arrayed or mapped in a way that might suggest deeper psychological relations between them. DAS, LAS, and CAS and other such programs all start with some kind of instructional task as given in the existing task taxonomies of traditional instructional domains. The tasks were chosen because they represent targets of difficulty in that domain. Each approach then

redesigns the task to improve instruction and assessment. Presumably, as work continues in each of these directions, adjacent tasks in pertinent domains might be encompassed and a network of tasks might be built up with instructional steps suggested between them. So, for example, work on LAS might expand to suggest how students with a certain pattern of performance in arithmetic word problems should be instructed to ease transition to parallel algebra word problems, or to figural arithmetic or geometric problems. Or, CAS might continue to expand not only to other job activities but to adjacent concepts in electricity and connections with general physics as well as with practical electronics. It is notable that related work on CAS for electronics has begun to identify the qualitative mental model progressions that lead toward expertise in this field (Frederiksen & White, 1990; White & Frederiksen, 1986). It is also especially important that continuing work on DAS, with its emphasis on recapitulating conceptual development, has now made significant progress in identifying some of the deep structural progressions involved in middle-level science and mathematics (Case & Sandieson, 1988).

These approaches can all be seen as "bottom-up" development of tasks for SIIA. In contrast, LPS is a "top-down" approach of a certain sort. It uses achievement test items, or the lists of pieces and parts of knowledge they reference, as a representative sample of tasks from a domain taxonomy to construct a molar network of performance tasks for that whole domain, based on psychometric considerations. Then it proceeds through analysis of those tasks down to more detailed cognitive process understanding of task performance to reach instructional and assessment designs for the intermediate steps. Unfortunately, if the LPS approach begins with achievement items not designed to reflect higher or deeper psychological goals of instruction, or targets of particular difficulty, it may well not reach

them. There is no necessity that the LPS approach start with conventional achievement test items; but it does start with performance tasks at a level that makes them candidates for use in conventional assessments.

But the two approaches, top-down and bottom-up, can be made complementary. An LPS map for a given domain might provide a network into which particular DAS or LAS or even CAS productions could be plugged, as noted earlier. If one considers any starting network for a domain, based on LPS psychometric scaling of performance tests or tasks for example, then the tasks studied from DAS, LAS, or CAS perspectives in that domain, can be located appropriately in that network. Empirical evaluations of network connections would then help show how the different approaches relate to one another, in a particular instructional design as well as for theoretical purposes.

From goal taxonomies to domain topographies. Another way to think about what is needed is to recognize that traditional approaches to both instruction and assessment are based on long established taxonomies of educational objectives. These may have logical or rational justification but they have little or no empirical or substantive theoretical basis. Such logical taxonomies need to be transformed into psychological topographies of instructional domains.

The attempt to build taxonomy in a scientific domain can be a useful exercise. It can sharpen distinctions, identify gaps in present knowledge, and systematize patterns of relationships among parts of a domain. In at least some natural sciences, such as biology or chemistry, building taxonomy has been an indispensable early step. The naturalist collects specimens of butterflies and pins them to a board. As the collection grows, rearrangements that systematize observed patterns of similarities and dissimilarities are

achieved. The patterns and classes can be used as a guide for further field research, as well as a starting point for analytic research on underlying principles and mechanisms.

The factor analytic tradition in differential psychology can be seen as an attempt to establish an empirical taxonomy of human abilities--a set of classification principles for types of ability tests (Vernon, 1950) and a search model for hypothesizing as yet unrecognized ability distinctions (Guilford, 1967). Some leaders of this tradition (notably Carroll, 1976; Thurstone, 1947) properly saw the result as only a first step, to be followed as experimental analysis of underlying processes within and across ability categories. There have also been attempts at provisional taxonomies for types of learning tasks (Gagné, 1965; Kyllonen & Shute, 1989; Melton, 1964; Merrill & Boutwell, 1973; Ryans, 1963) and for kinds of problem solving tasks (Greeno, 1978; Greeno & Simon, 1988; Sternberg, 1982). These in turn suggest cognitive structural and processing properties and distinctions that deserve further attention.

Unfortunately, premature or uncritical acceptance of a particular taxonomy can be quite misleading, even counterproductive in the long run, especially when apparent regularity is achieved by forcing nature onto a procrustean bed. Neat rows and columns can imply qualitative discontinuities where nature is continuous, and vice versa. A wholly different, more irregular structure may better represent the underlying psychological topography. In the domain of human ability correlations, for example, Guilford's structure has little or no empirical support as a taxonomy, even though the contents of some of its cells may be empirically justifiable (Cronbach & Snow, 1977). Snow, Kyllonen, and Marshalek (1984) used less restrictive methods to reach a radex structure for ability and learning task correlations; although obviously still crude, this structure seems to provide a

more useful topography or map of the empirical evidence on individual differences in abilities and learning. The points in the map represent concrete tasks, not abstract factors, even though constellations of points also represent hypothetical factor constructs; the tasks and the psychological distances between them can be investigated analytically to identify underlying structures, processes, continuities and discontinuities in the domain topography. And the map can be extended by further evidence without major distortion (see e.g., Ackerman, 1989).

By analogy, it can be argued that the study of instructional domains has suffered from a similar taxonomic problem and needs a similar transformation to some kind of psychological topographic representation. Indeed, the problem for instructional and assessment design is more severe, because its taxonomies were never intended to summarize empirical evidence. Increasingly over recent decades, the design of instruction, particularly by teachers in schools, has been guided either by rational taxonomies of educational objectives (e.g., Bloom, Englehart, Furst, Hill, & Krathwohl, 1956; Krathwohl, Bloom, & Masia, 1964) that assume generality across instructional domains, or by derivative or partial content classifications based on the general taxonomies but specialized for a particular use. Instructional content and task choices and also teacher-made tests seem to be governed, at least in significant part, by taxonomic distinctions of this sort. The design of standardized educational achievement tests has similarly been based on general, or at least partial, taxonomies. In the typical usage here, a content-by-process table is worked out in cooperation with instructional domain specialists. Weights are given to cells in the table to reflect relative importance. Then item writers construct test items for the cells in numbers corresponding to the weights. Some taxonomic and test development procedures

can be substantially more elaborate, of course (see Bloom, Hastings, & Madaus, 1971; Millman & Greene, 1989). For example, the Advanced Placement Test Program (College Board, 1988) uses committees of subject-matter experts and teachers at both high school and college levels, surveys of courses and textbooks at both levels, several question formats, and considerable pretesting, to arrive at its consensus products.

For some instructional purposes, such taxonomies may have important uses and no seriously detrimental effects. Good teachers presumably bring their own substantial pedagogical and content knowledge, and classroom experience, to bear on their instructional decisions and test designs; hopefully, they do not follow taxonomic specifications blindly. But there are teachers, unfortunately, who do see their task as simply covering the given categories. As used by many test developers, moreover, there is further cause for concern. The classification tables are well intentioned, but they are armchair constructions. The professional item writers are undeniably masters of their art, but their art is in writing items to fit such tables. The empirical analyses of the resulting items come late in the test development process, in the form of psychometric indices and expert judgments. The enterprise is usually justified only on the basis of content validity arguments and sampling theory.

The resulting tests are distributed, used, and increasingly watched in school, district, state, and national comparisons. They thus drive the educational system at all levels. If there is a problem in this, and many think that there is (Frederiksen, 1984), the problem starts with the taxonomies. Millman and Greene (1989, p. 350) are likely correct in stating that, "...the six general levels of Bloom's taxonomy of the cognitive domain...have productively guided many instructional test developers, especially by encouraging the

inclusion of relatively complex cognitive processes, in addition to the simpler ones." But the fact is that there is little if any substantive psychological theory or evidence to support the general taxonomy, much less the derivative content-by-process tables usually used by test developers. Nor is any psychological analysis conducted on the resulting items to demonstrate that they indeed reference particular cognitive structures and processes, much less "higher" or more "complex" ones. It seems clear that the time has come to require construct validation of educational assessments and the goal taxonomies on which they are built. At least to a limited extent, all four example systems can be seen as steps in this direction.

To understand the instructional psychology of a school subject-matter domain means to build a theory of learning, problem solving, thinking, and understanding in that domain. This is not a new ide (Judd, 1915, 1936), but it is an idea given impetus by the emergence of new theory and method in cognitive psychology. To advance toward psychological theories of instructional domains one must first reject generalized armchair taxonomies. Analyses of curriculum plans, traditional learning tasks, and expert teaching and learning in the domain can then guide the construction of a domain topography--a psychological landscape of the key concepts, procedures, structures and difficulties to be mastered, and the typical starting and ending points of successful and unsuccessful students. Such a map would have major gaps at first, because psychological analysis has focused only on individual school learning tasks to date, not on configurations or sequences of tasks. But such a map, however crude, seems a far better place for research and development on SIIA to begin than would be taxonomic abstractions promulgated by distant committees. Some such taxonomies may be useful in whole or in part. But that is a conclusion to reach late

in the progress of validation research on SIIA, not before it.

What is a domain topography? We see it as much like a combination of the learning progress maps envisioned for LPS and the goal lattices used in Lesgold's CAS, though not at first consisting only of reference tasks and subgoals. It will at first be a pinboard for all kinds of psychological and curricular "butterflies", including theoretical notions, empirical findings, and instructional task evaluations, gathered from whatever available sources, including old taxonomies and expert viewpoints. There will be no straightforward methods for studying psychological distances on this board as there are for ability test correlations. In many instructional domains, however, there are achievement test items and exercises with difficulty indices and intercorrelations based on large samples. Top-down analysis can at least begin with these. But bottom-up analysis must proceed in parallel. As noted previously, in some domains such as mathematics and physics there are rather many learning tasks and problem types that already have been subjected to isolated cognitive analyses. There is also the wisdom of teachers and other experts about which concepts and procedures students find most or least difficult or confusing. And there are curriculum theories that identify central conceptual structures and goals in a domain. Beyond the ETC approach, however, these analyses will need to address assessment as well as instructional design of tasks, and relations between tasks, as well as the psychology of each.

We choose to refer to topography as beyond taxonomy because we think that psychological depth as well as breadth must be captured in the representation of instructional domains. Difficulty is part of psychological depth for any given task. But so also is conceptual complexity, which need not be equivalent to difficulty. Complexity in

turn has many definitions--the number of components of knowledge and cognitive processing involved, the intricacy of connections in a knowledge network, the degree to which flexible adaptation of cognitive structure during performance is required. An important part of the research agenda has to be concerned with this breadth versus depth issue in the definition of a domain, and with the varying definitions of depth.

A topographic image. Figure 1 suggests schematically what such a topographic map might be like in the early stages of SIIA development. Imagine four related instructional end goals; A, B, C, and D. These might correspond to the expert viewpoints Lesgold uses to organize his goal lattice layer for a course. Each is reached by traversing, in one direction or another, a series of lessons that form a complete, coherent, and consistent course in Lesgold's terms. Since each lesson is associated with one or more performance tasks, there is also a student performance lattice. In LPS terminology, these tasks form a network of reference tasks. They may be arrayed along instructional and psychological difficulty gradients, as in the LPS approach. They may also be located at points of particular difficulty in the curriculum. But difficulty levels need not be equal across goals (see the BC or CD column comparison) and there may be differing numbers of tasks for different goals (see the AB or CD column comparison). Also, in keeping with the CAS approach, there are alternate routes that different students might take to the end goals (see especially the CD part of the map), and particular reference tasks might be adapted to fit the capabilities of different students at different points in time during instruction. To keep Figure 1 simple, we have not shown all possible connections between all tasks; nor have we indicated the varieties of instruction that might appear on the routes between tasks.

Insert Figure 1 about here

Both the CAS and LPS approaches assume some kind of hierarchical organization of subgoals, and associated instructional and reference tasks, and Figure 1 can be seen as representing this kind of structure. But the simple lattice need not imply a particular lock-step hierarchy of subgoals for all students to follow. The numbered boxes can represent instructional and assessment steps along a developmental sequence consistent with the DAS approach. The levels and columns can represent degrees and kinds of transfer as in the LAS approach. Also, because DAS and LAS address a fine grain analysis of performance, one can imagine one reference task (i.e., one numbered box in Figure 1) representing a whole DAS or LAS unit. The lattice then represents a network of such units covering a course of instruction.

The lattice concept also fits other kinds of tasks as well. Suppose a computer simulation such as Smithtown is inserted as a "high-level" task embodying goals A, B, C, and D. Students begin by exploring this simulation. As evidence accumulates that students are understanding particular concepts and procedures, the "lower-level" tasks are automatically checked off as accomplished. When evidence indicates that a student has failed to master some component, however, the appropriate subordinate reference tasks can be brought into play. As another example, suppose the numbered tasks under any one goal represent a series of student writing assignments or artistic productions, with teacher coaching in between; work on socalled "authentic" assessment uses such series (see, e.g., Gardner & Hatch, 1989; Gitomer, 1989). Each provides an opportunity for student

development and for exhibiting that development. Teacher critique provides the assessment, and directions for further steps. In short, the lattice can represent qualitative or quantitative stages in student development in a domain, not just hierarchically arranged instruction.

The inset in Figure 1 shows in more detail some of the instructional decisions that need to be considered with respect to a particular reference task B4. Performance on this task might indicate that vertical movement to B5 is the next best step for a particular student. Or, horizontal movement to A4 or C4 might be more appropriate. A combination of vertical and horizontal movement to A5 is also possible. In the main figure, this possibility is shown only for moves between D and C. Performance might also indicate that a particular form of tutoring or coached practice is needed on B4 before other instructional moves are considered. Or, it might indicate that a return to earlier steps is needed to remediate enduring difficulties. Along any line, as noted, other instructional interventions would occur between reference tasks. The map should be seen as superimposed on an instructional terrain, not as substituting for it. And the instructional terrain as presently taught may be quite rough, with many peaks and valleys, steep difficulty gradients, and canyons or gaps. Imagine raw mountainsides, not carefully groomed terraces.

It follows that psychological analysis of the instructional demands and opportunities, and also of the reference tasks themselves, will likely make the map much more irregular than that shown. Indeed, as such analysis proceeds the sources of difficulty in different parts of the map should come to be better understood. Difficulty becomes associated with alternative moves between reference tasks, not just with the tasks themselves, so psychological distances between different tasks may be seen to be quite different. This will

likely suggest the need for revised or different intermediate reference tasks in some regions, the redesign of instruction in others, and key changes in routes for learners differing in various prerequisites.

With continuing work, one can imagine that the network of reference tasks and instructional steps can be smoothed to overcome difficulties in the instructional domain as previously taught. To continue the metaphor, such a learning progress map should help turn a raw instructional mountainside into groomed terraces, more easily negotiated by both learners and teachers. However, such research may also suggest radical revisions in the map, and in the way instruction is organized in a domain. Suppose the bottom-up and top-down structuring of a domain fail to "meet" coherently, or otherwise suggest very different views of the instructional topography. Or suppose work on a particular reference task--a computer simulation of some sort, for example--suggests a radically different goal structure for a domain, or suggests different structures for different students, or teachers. Continuing research will need to examine the possibilities of alternative networks, and different maps for different purposes. Again, a provisional map is a framework for experimental teaching and research, not a grid designed for lock-step instruction.

Instructional domains do seem to have a "natural" topographic character, however, and each domain is to some degree unique in this character. The viewpoints of teachers and other domain experts, including textbook authors, combined with developmental theory and design (as in DAS), learning and transfer theory and design (as in LAS and CAS), and the psychometric results of broad test surveys, seems to provide the best starting points for LPS-style maps in each domain of concern. Continuing work can then aim to improve each topography wherever possible and to span or circumvent its inherent difficulties where they

cannot be changed.

The problem of scale. Locating reference tasks in a network is one aspect of what we think of as the "grain-size" problem, or the scale of the map. Even with a network of tasks that captures all the subgoals following Lesgold's criteria, for example, the psychological distances between subgoals or tasks may be great, from the learner's viewpoint if not from the teacher's or various experts'. But another important aspect of scale has to do with the graininess of diagnosis and instructional decision-making. In other words, instructional assessments are needed at different levels. To aid curriculum planning or school policy making, as well as to chart student progress in a domain, assessment needs to describe monthly or yearly student development; of our examples, only LPS seems to address this level. To assist teacher instructional planning and adaptation at the classroom level, assessment needs to focus on daily or weekly teaching and learning activities; LPS, DAS, and LAS address this level. But assessments to guide microadaptation of instruction, as in CAS, have to be focused at the level of transitions over seconds or minutes. One level of assessment may well not aggregate to another, or reduce to another. The two aspects of map scale are related; the distance between reference tasks, or the amount of intervening instruction, both dictates and depends upon the graininess of the diagnosis provided by the reference tasks.

At the finest grain level, researchers working on CAS think of this problem as the "learnable unit" (Wenger, 1987). They design instruction in assimilable increments. The size of these increments is limited by the learner's working memory capacity, by how well developed the cognitive structure already is when a unit of knowledge is incorporated into it, how coherent that unit is with existing structure, etc. There are also other possible

cognitive and technical constraints (e.g., only one disjunction, or branch of a decision point in procedural learning, should be presented in one lesson; a learnable unit should not exceed one frame on the computer screen). But the question takes on its other aspect when diagnostic assessment is considered. Burton (1982) applied a criterion of diagnostic discernability, to define the unit of knowledge as the smallest that can be individually mislearned, and detected as such. Complete knowledge or skill, then, is the avoidance of all possible errors on all possible problems.

A reference task (or series) designed as in CAS would need to grapple with this problem on a very fine-grain scale. At the classroom teaching or school curriculum levels of scale stand the pretest-instruction-posttest designs of conventional instructional evaluations; only a slightly finer grain is used in programs such as Individually Prescribed Instruction. For most SIIA, the aim is presumably to provide diagnosis on a scale somewhere between that of computerized tutors and weekly or monthly classroom tests. Some important points from the research on these two extremes are to be noted, however (see Lesgold, 1988, and Wenger, 1987).

First, diagnostic purposes place constraints on the definition of subgoals in the goal lattice structure. The subgoals need to be cast on a scale that allows important sources of error and misconception to be detected and also on a scale that effectively guides further instructional steps. Yet they also need to be on a scale that makes them meaningfully communicable. These scales may not have the same grain, and aggregation from one to another may not be possible, as noted earlier.

Second, diagnosis seems to require a relatively fine scale, also, to detect learner misperceptions that may not be explicit in the goal structure. This is what prompted

Burton's (1982) point about diagnostic discernability: in his work, subtraction problems of the form m-o or m-m were perceived very differently by learners than was the general case m-n, even though the goal structure did not distinguish them.

Finally, at any degree of grain size, there is a concept of confidence intervals to be developed; these have a substantive not just a psychometric character. To decide which of several next instructional steps to take, or which of several possible difficulties in learning occurred, the assessment needs to provide a confidence band for different next step decisions. Consider Figure 1 again. Each arrow entering or leaving reference task B4 has an associated confidence band that depends on the validity and reliability of the diagnostic assessment with respect to that decision. And we may want to require much greater confidence for some decisions than others. For example, the choice between moving a student to A4, A5, or B5 may not be critical if the consequences of a nonoptimal decision are not adverse or irreversible. However, deciding between any of these advances on the one hand, and further coached practice on the other, or between any of these and returning a student for remediation, may require high confidence (i.e., nonoverlapping confidence bands). In general, the more likely it is that negative cognitive or motivational consequences will derive from a diagnostic mistake, the more confident one must be in the instructional decision taken.

To emphasize the scale and confidence band problem, Lesgold (1988, p. 118) formulated "two fundamental laws of instruction": "Not everyone who passes a test on a topic knows what appears to have been tested" and "Not everyone who fails a test on a topic lacks the knowledge that appears to have been tested." In other words, each reference task, and the distances between reference tasks in a map, need to be designed

in a way that minimizes these sources of diagnostic error. Also, in considering psychological distances between reference tasks in the map, a further "law of instruction" (formulated by Snow, 1972, as a lesson from ATI research) needs to be considered: "No matter how you try to make instruction better for someone, you will make it worse for someone else." In other words, the instruction that intervenes between two reference tasks may be optimal for some learners, but it may create cognitive or motivational problems for other learners, particularly if it is disjunctive with what they have already come to know or do well. Reference tasks need to be designed to detect these problems for each instructional route that connects to its node.

In short, the scale problem in a learning progress map of goals in a domain has important instructional and diagnostic aspects. In the early stages of SIIA development in a domain, it is not likely that much will be known about this problem.

Reference Tasks and Teacher Assessments

We have already noted that the kinds of tasks designed for DAS, LAS, and CAS are themselves candidate reference tasks within larger curriculum domains. So too are computer simulations such as Smithtown. The computer-based tasks also produce collateral indicators of performance differences. LPS starts from selected test tasks such as those used in NAEP, another important source. However, there are three further sources, each related to one of the above.

Experimental cognitive tasks. First, there are myriad cognitive psychology experiments designed to exhibit aspects of cognitive structure and processing. Some use instruction-relevant tasks and some do not, but even these sometimes suggest instruction-relevant analogs. The value of these tasks is in the attempt to distinguish deep cognitive

functions sharply. Even with analog reference tasks one can hope to make the same diagnostic distinctions. Many of these studies also develop scoring formulae for think-aloud protocols or retrospective interviews, and these could be included in reference task design. Although much of this work has been small-scale and piecemeal, current research is branching out from circumscribed examples to cover larger ranges of instructional tasks in mathematics, natural science, and language skills, and also increasingly in other school curriculum domains, including social science, history, and art. A systematic catalogue of these tasks would be an extremely useful start for reference task development.

Achievement tests. A second source, as suggested by the LPS project, is the vast population of achievement test items. These also could be catalogued, along with their known psychometric properties. There have been attempts to build algorithms for generating test items that include systematic manipulation of cognitive aspects of items (e.g., Bormuth, 1970; Butterfield, Nielsen, Tangen, & Richardson, 1985). There are also new procedures for judging and coding the cognitive demands of existing items (Carroll, 1976; Emmerich, 1989). A catalogue of items indicating both psychometric and cognitive properties could be of great help in initial mappings of reference task networks.

A new area of research addresses various possible constructed response formats for achievement tests. It recognizes a continuum of performance tests ranging from authentic instructional tasks to artificial multiple choice tests. The various problem types along this continuum each present different costs and benefits in their provision of assessment information about student performance. The view that multiple choice problems at one end of the continuum are "bad" and that production or constructed response problems, at the other extreme, are "good" is an oversimplification and distortion.

Bennett (in press) describes one program of research on constructed response items, which he defines as "any task for which the space of [potential] responses is not limited to a small set of presented options" (p. 1). In this format, students must actually construct a response, more or less as they do in activities readily encountered in classroom settings. Hopefully, this format is more likely to engage and promote deeper thinking, reasoning, and understanding, and thereby exhibit important learner strengths and weaknesses.

However, there are theoretical and practical issues that must be considered with the constructed response format. Multiple choice tests generally present students with a broad range of content in a short period of time. Constructed response formats offer partial credit and thus more detailed diagnosis. Bennett (in press) notes that the distinction often made between the two extremes is that multiple choice items often are deemed irrelevant, whereas constructed response tasks represent more directly the real-world performance of interest. However, constructed response tasks often display lower reliability and validity, require complicated scoring procedures, and involve difficulties with interrater agreement. These issues make the simple good/bad dichotomy a misrepresentation. There are tradeoffs. For present purposes, constructed response items of the sorts now being studied represent a further source of ideas for reference tasks.

Teachers and test exercises. Teacher observations of student performances and judgments of student productions, as well as the questions, exercises, and "testlets" routinely used in teacher discourse or embedded in text, provide a third source. Especially useful may be the classroom exercises now being devised to study and promote deep student understanding in particular school domains (e.g., Lampert, 1986, 1990; Palincsar & Brown, 1984). Here, also, a catalogue of the means by which teachers make face-to-face

assessments of student progress and judgments about student productions would provide a starting point for reference task construction. Although teachers would appear to be regularly engaged in this kind of moment-to-moment diagnosis, research on teacher's interactive decision making seems not yet to have focused on individualized assessment decisions explicitly (see Clark & Peterson, 1986).

Reference task development and probe tasks. Although there is a developing cognitive psychology of reference tasks, there is rather little research to rely on regarding the design and evaluation of reference tasks within a learning progress map for a course. In particular, the transitions between references tasks will need to be explored in task design because these transitions are the key to instruction.

An initial step worth considering would define two classes of tasks for instructional assessment in a domain: reference tasks for which there is already some degree of understanding of their cognitive learning and performance properties; and probe tasks that appear well chosen on substantive, logical, or statistical grounds, but for which there has been no cognitive analysis. The probe tasks, however, should be chosen primarily for their value as vehicles by which to explore the topography of the cognitive "ground" around and between the reference tasks. Whether reference and probe tasks scale statistically along smooth gradients of difficulty in the domain is of secondary importance. Indeed, we think existing achievement tests are useful as starting points precisely because their items cover substantial ranges and types of difficulty. But such items should be viewed as probe tasks, not yet reference tasks, and they should probably be used in constructed-response not multiple-choice format, if they are to be effective probes. Similarly, teacher and text exercises are initially probe tasks.

Probe tasks could signal problems to be examined more fully in reference task analysis and could also identify places or topics for which additional reference task construction was needed. Evidence from teaching following a reference task could help evaluate that task as well as help teachers learn to use its data for diagnostic purposes. Suggestions for adaptive teaching between reference tasks could be obtained. Evidence from instruction leading into a reference task could also help guide their redesign to improve diagnosis and enlarge the learning progress map. A related function for probe tasks before or after a reference task would be to check and hopefully increase confidence in instructional decisions based on reference task performance. They would provide follow up evidence that diagnoses and prescribed instructional routes were indeed having expected effects.

Still other functions are possible for probe tasks. In the instructional coaching and practice segments of reference tasks, but also in teaching between tasks, there need to be guided series of interactions between learner and teacher that have a combined diagnostic and instructional function, as suggested by the hint structures and scripted tutoring of LAS. As previously described, there would be a hierarchically organized sequence of helps brought into play as the learner meets difficulties. The hints are used by the teacher in a prescribed order so that, in addition to providing instructional help, they also indicate the type and depth of help needed by a learner. The teacher probes could also address transfer. An array of such tasks could provide a wide variety of practice across related problem situations. Perhaps a degree of learner choice could also be allowed in this array. The "probe" aspect of the array, however, is to permit teacher checks on flexibility and transfer of learning as part of diagnosis. These could also be organized in a way that lets

learner choice, as well as degree of learner success or failure, signify features of knowledge structure and process of use to the teacher. In short, we think that a bank of probe tasks associated with each reference task, to be drawn on by teachers and learners during intervening instruction, would be an important auxiliary to a learning progress map.

Teacher-friendly design. At several points in the preceding discussion, one or another role for teachers was suggested, and such notes appear in following sections as well. But it is important here to emphasize the basic need for research designed to understand and shape the interface between teachers and SIIA in all its facets. We believe no progress can be made in educational practice related to assessment without it.

None of our example systems address this need explicitly, although all refer to teacher roles, and LPS development plans include studies of teacher preparation and use of the system in later stages. Both DAS and LAS provide detailed prescriptions for teachers to follow with individual students or small homogeneous groups. But the working conditions of teachers in most classrooms most of the time preclude routine use of such procedures. Most work on CAS essentially ignores the teacher interface.

Research on all four systems can produce principles as well as procedures that some teachers might use some of the time. These might also be adapted for student use in small group, cooperative, and reciprocal teaching (see e.g., Palincsar & Brown, 1984; Webb, 1982). Furthermore, all four systems might be used profitably as demonstration vehicles for discussions of learning and assessment in teacher education programs. There is at least one demonstration that teacher expectations for slow learning children have been enhanced by having teachers observe those children performing in LAS conditions (Vye, Burns, Declos, & Bransford, 1987).

But the core problem is one of designing SIIA for effective use by teachers and for effective placement into the context of ongoing classroom instruction. The research needs to obtain teacher input early in the process. It needs to seek an understanding of how teachers make assessment judgments without help at present, and how teacher and system assessments can be made complementary. In effect, any SIIA is an attempt to do what a good teacher could do with a class size of 5 students instead of the typical 25 or 30. Viewed in this way, SIIA design needs to be carefully articulated with important aspects of the teacher role change that is forced by this difference in student numbers. Closely related are questions concerning how teachers can give input to SIIA operation and how output from the system can be formulated for optimum teacher use. In short, effective SIIA will have to be teacher-friendly, and this requires research on system-teacher interaction in early stages as well as late stages of system development.

The Nature of Learning from Instruction

Any assessment design has to include assumptions about the nature of learning from instruction because it has to be geared to detect and characterize it. The four research programs we use as examples, and the bodies of other research related to each, are part of the tremendous expansion of instructional psychology over the past two decades. As a result, there are many new conceptions about the psychological structures and processes involved in learning and development in relation to instruction, and of individual differences in aptitude and achievement as well (Glaser & Bassok, 1989). The work is also now engaging a larger range of the critical issues that will need to be understood for future improvement of instruction and assessment. These issues concern: (a) the many kinds of cognitive representations, skills, and strategies that might be expected in school learning,

the degree to which they are general or domain- or situation specific, and the means by which they can be diagnosed; (b) the kinds of human capacity limitations that influence school learning, such as working memory load limits, and the means by which they can be minimized; and (c) the nature of expertise and the alternative possible forms of novice to expert progression. This section examines these issues.

As with all previous approaches, however, cognitive instructional psychology has also until very recently begged two further questions important for SIIA: (d) it implies that instruction must be adapted to individual differences in learning, without specifying how this is to be done; and (e) it imposes a cognitive analysis of learning and transfer, without addressing how conative or affective aspects of performance can be brought into the picture. This section also takes up these questions.

Multiple cognitive structures and processes. To consider the kinds of mental structures, processes, and capacities that may need to be diagnosed in learning, we here review briefly some of the developments in cognitive psychology. This also will help us consider the generality-specificity issue in school learning assessment.

Through the 1960's and 1970's, research aimed at building cognitive theory focused on the basic processes of attention, memory, reasoning, and problem solving. Relatively able young adults were studied performing relatively simple, well-specified game-like or puzzle-like tasks that required little or no background knowledge. Some of these tasks were specially designed to test theoretical distinctions. Others corresponded closely to the kinds of tasks found in tests of fluid intelligence, or abstract reasoning, or various memory or perceptual abilities. Mixtures of experimental analysis and simulation modeling produced cognitive models of many particular tasks. In addition, there developed some

theoretically useful typologies of such tasks and important hypotheses about the kinds of cognitive structures and processes that might be basic to performance in all tasks within a class, and perhaps also across classes. The work thus identified some of the possible general skills able students use to generate representations, plans, transformations, and evaluations in almost any novel task for which they have not been trained. It also showed some of the limits on such processing that are imposed by cognitive system capacities.

From the research on problem solving, for example, a candidate general learning skill might be using problem instructions to construct a representation of a problem space, with appropriate goals, operators, and constraints. The efficient use of means-end analysis within such a space might be another. So might the use of planning, to reduce the space that must be traversed. Still others might be the component processes involved in apprehending patterns of relations among parts of a problem and in rule induction once a pattern is perceived.

We could construct a list of problem solving abilities and processes such as these from reviews of this literature (see Greeno & Simon, 1988; Simon, 1976, 1978; Sternberg, 1982, 1985). Greeno's (1978) classification of problem types and their associated abilities is especially useful for this purpose. The result would be a set of abstract thinking and reasoning skills hypothesized as generally relevant to school learning whenever instruction was novel, or incomplete enough to require learners to problem-solve across the gap, or when for any other reason learners had to fall back on these skills in the absence of content knowledge. One could then develop measures of each such skill for inclusion in SIIA, along with training exercises to promote their improvement. Some current research programs have precisely this goal--to diagnose and train the constituents of fluid

intelligence directly (see e.g., Baron & Sternberg, 1987; Chipman, Segal, & Glaser, 1985; Detterman & Sternberg, 1982; Lidz, 1987; Nickerson, Perkins, & Smith, 1985; Segal, Chipman, & Glaser, 1985; Sternberg 1986). This is one aim of the research on LAS.

On the other hand, when two abstract problem solving tasks of the same typological class are compared, they often seem to demand somewhat different configurations of constituent skills even if the same skills are nominally involved in each. When tasks from different classes are compared, the configuration of skills appears further specialized and somewhat different skills seem to pop in and out, even when total scores for the two tasks are highly correlated across persons. Thus, a further hypothesis is that it is not only the possession of all these skills but also their flexible assembly into particular organizations or strategies for particular tasks that is crucial to successful performance (Brown, 1978; Snow 1981; Snow & Lohman, 1984). This is an important emphasis for both LAS and DAS.

There is also evidence that effective problem solvers learn within a task to assemble a good strategy by shifting and tuning as experience accumulates; the evidence suggests that they do so in keeping with their prior ability profile, as it applies to particular tasks (Kyllonen et al., 1984), and that some important kinds of shifting and tuning within a task can be captured only by forms of cognitive analysis that descend to the level of eye movement tracking (Bethell-Fox, Lohman, & Snow, 1984). We are left with a structure of abstract thinking and reasoning skills which may be hypothesized as general and transferable to much of school learning. Yet we are also left with the intriguing but difficult hypothesis that what is left out of such a structure is the crucial ability to assemble and reassemble such skills into flexible strategies during learning--to construct and adapt one's performance to each particular situation as it occurs.

A related picture comes from research on memory and attention. Here there is the general concept of a limited capacity working memory, as contrasted with a permanent and unlimited long term storage, and the related concept of a limit to attentional resource allocation. These limits apply across many performance situations. For example, developmental theorists such as those working on DAS posit a limit on the number of schemas that can be active in working memory at any one time. Task requirements that exceed this limit cannot be successfully met. The limit varies across persons as well as within persons across age. Within an age, other general skills or capacity limits are hypothesized to influence learning. Speed of selective reaction to stimulus displays and speed of access and search of semantic memory are examples of measurable and presumably general characteristics of the cognitive system that do not vary across task situations appreciably. Ability to function efficiently under dual attentional demands may be a general skill as well (Hunt & Lansman, 1982). General attentional heuristics may also operate (Ohlsson, 1984a). And automatization is a general process that relieves attention demands in any task that becomes consistent across trials (Ackerman, 1989).

There are also methods of memory enhancement that appear general. A long list of learning and memory strategies has been developed, including mnemonic devices, uses of imagery, strategies for enhancing key word identification, schematic mapping, conceptual organization, rehearsal, and also metacognitive awareness and control of memory function. All these techniques appear to be trainable and widely applicable across many different learning situations (Baddeley, 1982; Brown, 1978; Norman, 1980, 1982; O'Neil, 1978; O'Neil & Spielberger, 1979; Weinstein, Goetz, & Alexander, 1988). So again, one could choose to equip SIIA for general attention and memory assessment and direct training of the

constituent skills.

Beneath or beyond these general abilities, however, there must be considerable specialization of knowledge and skill. At the level of thinking and reasoning with relevant prior knowledge in tasks that have become to some degree familiar, or are perceived to be so, it may be the specialized assemblage of situation- or domain-specific knowledge and skills that primarily governs performance (Glaser, 1984; Glaser & Bassok, 1989; Greeno & Simon, 1988). This view arises from a more recent phase in cognitive psychology through the 1980's, wherein research on learning (Greeno, 1980) and development (Case, 1985; Siegler, 1986) has increasingly addressed the kinds of complex learning and problem solving tasks found in school instruction, and in tests of crystallized intelligence and achievement. The aim is to understand the knowledge structures and processes required in particular school performance domains and the manner in which these are acquired during instruction. The research has also widened the range of persons and thus of the ability and knowledge levels studied. Through this work, learning and development have increasingly come to be described in similar terms--as the replacement of primitive, limited, or naive cognitive structures by more advanced, complete, or functional structures tuned to the conditions of their use.

Beyond the general skills of able learners and problem solvers in novel tasks, then, this research shows the important uses of domain-specific knowledge and skill in familiar tasks, and the role of naive beliefs, misconceptions, and faulty reasoning in a domain. It begins to sketch the cognitive constituents of domain mastery, and provides guidelines about the kinds of knowledge structures and learning processes to look for in analyses of learning in particular domains. Still further development has come from research on

situated learning, which sees knowledge and skill acquisition as a specific person-situation interaction, highly dependent on the particular demands and affordances of each situation. The issue of general versus domain-specialized versus situation-specific knowledge and skill, and their integration in particular performances, will no doubt continue to fuel controversy and research. (For recent discussions see Brown, Collins, & Duguid, 1989a, 1989b; Glaser, 1984; Palincsar, 1989; Perkins & Salomon, 1989; Resnick, 1987; Wineberg, 1989.) The issue is important, because the various views suggest that the human cognitive system has several layers wherein rather different principles may operate. At one level are the multiple, general variations in processing skills and structures, and also in capacity limits, that transfer and apply across domains; they influence performance especially in novel situations and thus in early phases of learning. At other levels are the multiple, specialized skills and structures tuned to the conditions of particular domains and perhaps to specific processing situations; these presumably grow into prominence with increasing familiarity with particular domains and situations, and thus influence later phases of learning. Within and between levels, however, it may be the flexible assembly of these capabilities in response to situational demands, affordances, and changes therein, not just their possession, that is crucial.

We need not try to predict here just what theory further research will suggest. But we do need to consider what these views about general and domain knowledge might lead assessment researchers to expect, at least in summary terms. In other words, what kinds of learning processes, phases, and structures should assessment research address in designing SIIA for any particular domain?

Phases and structures of learning. One useful summary view of cognitive learning

divides learning into three modes or phases: accretion, structuring, and tuning. These are not distinct stages; they may overlap, operate concurrently, and also recur at higher levels of learning. For the most part, we rely on the terminology of Norman (1982; see also Rumelhart & Norman, 1978). Since this view also applies to complex psychomotor skills, we also bring in ideas from Ackerman (1989) and Kanfer and Ackerman (in press) on cognitive and motivational aspects of skill acquisition, as well as Anderson's (1985) theory.

Accretion is the addition of new facts, concepts, or procedures to long-term memory via working memory. The addition may establish an entirely new region of knowledge, but usually new parts are also connected during accretion with parts of previously stored knowledge. The work in working memory is to make these connections, to elaborate, chunk, and personalize the new knowledge--in short, to make it meaningful. The process is often slow and error-prone, though it will be less so for persons with more relevant prior knowledge and ability. In skill acquisition, but also in most school learning, this phase would also include adding declarative knowledge of the goals, objectives, and task requirements, and constructing a cognitive representation of the task. The requisite general memory and reasoning processes will have to be engaged, and considerable attentional resources will need to be devoted in this phase. The motivation to engage in learning here, at least when there is little prior knowledge of the task, probably comes from distal sources: need for achievement, perceived utility of the knowledge in the future, general interest, and expectations of success.

Structuring is the formation or compilation of a new knowledge organization--an interrelated set of parts to form a new concept, or system of concepts, or integrated procedure to reach some goal. The new structure may be entirely new--a new region

chunked. Often, however, the new organization replaces or reformulates an old one, in whole or in part. If so, then the restructuring requires the unlearning of old relations, the jettisoning of some old parts, and thus much attention and effort. Motivation here come from proximal sources: the strength of intention-action commitment to the task and other self-regulatory skills. Given that all students arrive at a school learning task, or a reference task, with prior knowledge that is partially correct, partially faulty, and in some major ways incomplete and disorganized, the process of structuring and restructuring may be the crux of the instructional problem. Unfortunately, a large amount of prior research in instructional psychology seems to have assumed that accretion is the primary issue.

Tuning is the adjustment or adaptation of a knowledge structure to particular uses. Because learned structures may be overgeneralized and because new situations calling for their use almost always vary, mismatches of some deg. ·e will occur and fine tuning will continue to be required. Long hours of practice and expe. ·nce are needed, in the family of situations to which knowledge structures apply, to change mere knowledge into expert performance. Even after procedural skill is automatized, practice continues to improve, to specialize, and to situate the assembled ability. Motivational requirements here are probably a continuation of those instrumental in the structuring phase; proximal persistence and action control are the hallmarks. Eventually, the exercise of specialized expertise probably carries its own intrinsic motivation.

What kinds of knowledge structures might be formed by such learning, and what kinds of indicants of their characteristics might be observed? Again, Norman's (1982) language is a convenient starting point, which can then be elaborated (see also Mandler, 1984). It is recognized, of course, that these are hypothetical constructs; the claim is that

evidence exists to suggest that people behave as if they possessed these kinds of structures.

1. Semantic networks represent accepted facts, concepts, and events as nodes in a memory structure, and their interrelationships as arcs connecting the nodes. The arrangement may be hierarchical with class inclusion, as in biological classification systems, or involve broken or multiple interlaced hierarchies, or it may be simply associative. Other types of networks such as serial or matrix structures are also possible. Networks can also include if-then propositions, contradictions, exceptions, and other conditions. New knowledge can be deduced by reasoning through a network and then added to it.

2. Schemas are highly integrated or unitized structures that can include both a declarative knowledge network and procedural rules for its use in some domain. They thus provide inferences for that domain. But schemas can be quite general, and they can contain references to other schemas. Typically they organize spatiotemporal events, so there are event schemas, scene schemas, and story schemas. But there are also different schemas for arguments, information expositions, and other functions of text, and probably also for different classroom activities. A schema has slots into which information can be placed in a particular instantiation. Procedural rules are thought of as if-then or condition-action schemas--if a certain condition is satisfied, then perform this action. But there is also a distinction between goal-action schemas, where the condition is an intention to reach a specified goal, and a trigger-action schema, where the condition is the occurrence of some specified cue or triggering situation. There are thus schemas within schemas, and some theorists have regarded schemas as the basic building blocks of knowledge structure.

3. Scripts are a specialized form of schema distinguished by some theorists to represent and account for highly tuned, routinized, or stereotypical patterns in event

memory.

4. Prototypes are declarative or procedural structures built on a goodness-of-fit criterion to account for the fact that learners do not always operate as if their knowledge was organized in semantic network form. For example, the concept of "animal" includes "wolf", "person", "penguin", and "sponge", but these differ (in descending order) as to the degree to which they fit the prototypical or ideal concept of "animal". Knowledge about animals is likely to be orderly when it is about prototypical animals, such as wolves, but disorderly when it is about penguins and sponges. Presumably, a script is a procedural prototype.

5. Images are picturelike mental representations that allow contemplation of stimuli, sometimes quite complicated objects, events, and scenes, in their absence. Some learners generate and transform images to capture their understanding of complex phenomena and to aid memory. But particular images are subject to errors of omission and intrusion, just as are other knowledge structures, and their details can become distorted toward stereotypical or prototypical schemas (see Kosslyn, 1980).

6. Mental models are mental structures learners build to understand the functioning of some target system, usually a complex physical or natural phenomenon or working machine, but also possibly an abstract system such as a syllogism. Instead of running the machine or algorithm over, a mental model of it can be "run". Contingent actions can thus be planned ahead of time. Mental models may be thought of as a type of schema often involving imagery. They too are subject to distortions, incompleteness, confusions, and instabilities (see Gentner & Stevens, 1983; Johnson-Laird, 1983).

These sorts of knowledge structures are not only generated in the course of learning,

they are used to reason about what is being learned, to recall what has been learned, and to solve problems. Their effects are thus sometimes detectable in performance. Snow & Lohman (1989; see also Snow, 1989a) have discussed some of the measurement possibilities. For example, the degree to which a learner has restructured an instructional presentation, as opposed to memorizing it verbatim, may be seen in the degree to which information is paraphrased, rearranged, and chunked differently in a recitation or teach-back session. Size of chunk can also be inferred by using interresponse time intervals to segment the recitation into chunks and then counting idea units within chunks (see, e.g., Chi, 1978; Gray, 1982). Scoring templates for different kinds of structures can be imposed on protocols obtained in these teach-back sessions, or in interviews or problem solving think-aloud sessions. It is known that qualitatively different kinds of knowledge structuring can be detected in such protocols, for different persons but also for different instructional domains (Marton, 1983; Marton et al., 1984; Pask, 1976). Marshall (1988, 1990) has devised procedures for estimating the various nodes and arcs in mathematical knowledge schemas. It is also known that different kinds of semantic structures in memory tend to produce different patterns and sequences of responses on recall, and particular kinds of errors (Mandler, 1984). Presumably, performance tasks can be designed to elicit different kinds of errors from different structures. There are also a variety of word association, graphing, card sorting, interview, and questionnaire techniques that have been tried with some success (McKeachie, Pintrich, & Lin, 1986; Naveh, Benjamin, McKeachie, Lin, & Tucker, 1986; Pines, Novak, Posner, & Van Kirk, 1978). Further, a considerable amount of research on the training of learning strategies is at least suggestive of methods for assessing knowledge structuring and student approaches to it (O'Neil, 1978; O'Neil &

Spielberger, 1979; Weinstein et al., 1988).

Other indicants of knowledge and skill acquisition, however structured, can also be considered. For example, Heller and Greeno (1979) contrasted high and low skill in four aspects of word problem solving in physics and mathematics as shown in Table 2.

Insert Table 2 about here

By distinguishing these aspects, they could reach concrete descriptors for assessing extreme skill differences, though not intermediate levels between these. Based on their review, Snow and Lohman (1989) offered a list of six aspects of skilled performance that could be displayed and assessed once some degree of skill had been acquired: correctness of output (correct execution without consistent errors); conceptual foundation (understanding the principles underlying the skill as opposed to mindless execution); automatization (rapidity of execution with minimal demand on attention resources and minimal disruption by concurrent tasks); degree of composition (smoothness and speed of execution of multiple actions organized as a unit rather than as separate steps); generalization (range of performances to which skill readily transfers); and metacognition (appropriate action control and self regulation). And Norman (1982) provided another list of five aspects that distinguish expert performance from very good performance. In addition to automaticity and smoothness, as above, he also listed mental effort (decrease in effort and fatigue as skill increases), performance under stress (no deterioration of skill under stressful conditions), and point of view (the expert becomes part of the action system--the expert pilot just "flies", rather than "flies the plane").

It is clear that there are multiple indicators of declarative knowledge and procedural skill acquisition along the progression from novice to expert. These are candidates for use in assessment tasks. However, the actual design and validation of knowledge structure and skill assessments in instructional domains remains an uncharted new frontier for research. Our example systems offer a start, but leave many issues open.

There is a parallel between the argument about general versus domain-specialized versus situation-specific learning and arguments about the role of different kinds of abilities in perceptual-psychomotor skill acquisition. Ackerman's (1989) theory and evidence in the latter domain suggest three phases of skill acquisition that parallel Norman's (1982) phases for learning in general. In the first, or cognitive phase, the learner must build a cognitive representation of the novel task and its requirements, so individual differences in general intelligence and major content abilities, such as verbal or spatial abilities, play a predominant role. In the second, or associative phase, the learner compiles sequences of cognitive and motor processes into a production system for the task; general abilities become less relevant, while perceptual speed abilities come to predominate. In the third, autonomous phase, where automaticity is reached, cognitive abilities no longer limit performance; individual differences are a function of performance asymptotes determined by psychomotor abilities. General cognitive abilities will also be less involved in consistent tasks and more involved in complex, inconsistent tasks; they thus will reappear as important in transfer tasks. By analogy from Ackerman's theory, general and metacognitive skills and processes may be more involved in the early phases of accretion and structuring, domain-specialized skills and processes may be more involved in later phases of structuring, and tuning may then become increasingly situation specific and thus skill specific. General

skills and processes reappear as essential for transfer to new situations, however, because such situations are by definition inconsistent with the situations of original learning.

Alternative views of expertise. Further problems are associated with assessing progress toward expertise. There are at least two distinctly different views of what constitutes the development of expertise in a domain.

Glaser (1985) has presented the view deriving from research in cognitive science, including artificial intelligence, in which many kinds of experts, intermediate performers, and novices have been compared. The work covers empirical studies of reasoning and problem solving, as well as simulation modeling and the design of expert computer systems (see also Greeno & Simon, 1988). In general, it is assumed in this view that human intelligence is based on symbol processing and that expertise is built up by specializing abstract systems of rules for symbol processing in a domain. As Glaser (1985, p. 4) summarized it, "...novices' representations are organized around the literal objects and events given explicitly in a problem statement. Experts' knowledge...is organized around inferences about principles and abstractions that subsume these factors...In addition, experts' ... declarative information is tightly bound to conditions and procedures for its use. An intermediate novice may have sufficient knowledge about a problem situation, but lack knowledge of conditions of applicability of this knowledge." Glaser (1985) offered a set of 24 propositions as a detailed summary of this view. We provide the following condensed and paraphrased version, without noting each quoted line or phrase:

Expertise is developed over hundreds and thousands of hours of learning and experience, and continues to develop. Along the way, plateaus and non-monotonicities of development appear to indicate shifts in understanding and stabilizations of automaticity.

Experts and novices are similar in processing capacities; the difference is in how knowledge is structured for processing. In developing expertise, problem representation changes from surface representations to inferred problem descriptions, to principled, proceduralized categorizations. Expert schemas are like fast-access pattern recognitions that reduce processing load and the need for general heuristic skills; they have actionable meaning, bound to conditions of use. Experts are opportunistic planners, with fast access to multiple possible interpretations; novices are less flexible. Experts automate basic operations so that working memory is free as necessary; they show skilled self-regulation, in solution monitoring, attention allocation, and sensitivity to informative feedback. Experts are schema specialized, schema driven, goal driven, and efficient; they will think only as deeply as necessary and use general skills usually only in unfamiliar or ill-structured situations--they display superiority over novices in domain intelligence, not general intelligence. But expertise in some domains is more generalizable and some super experts generalize by mapping and analogy to other domains. Finally, general skills may be developed by this shifting among domains, so that cognitive processes become decontextualized.

It is worth a note in passing that at least one expert simulation model built in the symbol processing tradition seems to possess generalization and transfer capability. "FERMI" (Larkin, Reif, Carbonell, & Gugliotta, 1985) is a problem solving system that is not only expert in one domain but can also respond flexibly to novel situations from other domains. Designed initially to solve problems in the physics of fluid statics, it also appears able to transfer its expertise with minimal additions to solve problems in aspects of electronics and chemistry. It achieves its generality by encoding declarative and procedural knowledge separately in hierarchies that also distinguish domain specific and general

knowledge at different hierarchical levels.

However, an alternative view of the cognitive psychology of expertise has been put forth by Dreyfus and Dreyfus (1986). They criticize the artificial intelligence, symbol processing, computer model of expertise as fundamentally incorrect. Human experts are not at all like analytic engines, applying rules and making inferences with great speed and accuracy in a large, domain-specific knowledge base. Rather than learners beginning with specific cases and moving to more and more abstracted, sophisticated, and automatic rules, they argue that skill is acquired in the opposite direction--from abstract rules to specific cases. Thus, rule-based symbol manipulation models of expertise mislead us away from the study of intuitive expertise based on human experience.

The Dreyfus description of skill acquisition via instruction posits five stages through which a learner must progress; these are termed novice, advanced beginner, competent, proficient, and expert. Initially, instruction is usually designed to decompose tasks into context-free features and rules for operating on these features. The novice learns to recognize these facts and features and to follow the permissible rules. Performance is context-free because the key stimuli are isolated from their situation and no coherent sense of the overall task or goal is given. Performance is judged by how well and consistently the learner follows the rules. Then, with increasing experience with more and more concrete situations, the learner becomes an advanced beginner who recognizes meaningful elements in new situations and perceives similarities with previous examples. The learner thus comes to operate with rules referring both to context-free elements and also to situational elements. The latter are not easily captured in words and so are not easily explicated in instruction, but they are implicit in well chosen instructional examples. The Dreyfus

examples of this stage are the driver learning to use engine sounds (situational) as well as the speedometer (context-free) in rules about gear-shifting, the chess player learning to recognize overextended positions or strong pawn structures without precise context-free definitional rules, or the nurse learning from experience to use a patient's breathing sounds as well as formal measures to recognize or distinguish problems.

The third stage, competence, is marked by adopting, or being taught to use, hierarchical plans to organize patterns of experience in situations in order to identify what is important in a particular situation and to guide sequential decision-making accordingly. It is choosing a perspective, a plan, or a goal needed or demanded in a situation, based on a constellation of situational features, not on objective, context-free rules. Thus, it involves both necessity and uncertainty. There is also the emotional feeling of responsibility for the outcome, and vivid memories of resulting successes and failures as whole situations. The competent performer stops thinking of problem solving as a detached observer following rules, and starts thinking within a gripping holistic experience of a situation. The Dreyfus argument asserts that the research on cognitive information processing has shown that a theory based on physical symbol systems drawing inferences with features and rules is sufficient to account for intelligent human problem solving, but in no way shows that such a theory is necessary; much of human intelligent behavior seems not to be rule based, so perhaps human expert problem solving is also not so based. Especially in the highest two levels of their skill model, the rapid, fluid, involved character of performance seems quite dissimilar from the slow, detached careful reasoning of the problem solving process.

With long accumulation of experience in a family of situations, the proficient performer effortlessly sees similarities between past situations and the one currently faced,

and has a perspective that makes some situational features salient. These similarities and perspectives change gradually with experience as do the plans and expectations they trigger. The result is intuitive grasp of the present situation--a kind of "know-how" that has in the past been attributed only to feminine interpersonal skill and considered inferior to masculine rationality. But proficient performers still think analytically about what to do, given an intuitive understanding of the situation; they see what is needed but decide how to do it. The expert, however, has also discriminated classes of situations that share the same decision, action, or tactic. This allows response as well as understanding to be intuitive and immediate. These masters become one with the game; they do not see themselves as independent, manipulating pieces "out there." Expert drivers, pilots, and sailors become one with their crafts, as in the Norman (1982) example of tuning noted above. The Dreyfus argument is that expert doctors, nurses, business managers, and also physicists and mathematicians, do the same. With expertise comes fluid, flexible, integrated performance. This is not to say that experts do not deliberate. When time permits and outcomes are crucial, experts reflect on their intuitions, but this does not involve analytic problem solving.

These contrasting views of expertise leave us with three questions which we address below. The first concerns whether the theoretical disagreement is really irreconcilable. Secondly, even if so, one can ask whether the implications of the two views for instruction and assessment are fundamentally opposed. The third question is whether either view helps define the progression of levels of learning in a domain that need to be assessed.

The theoretical developments have not yet run their full course. The Dreyfus view says that orthodox AI pursuit of expert systems is doomed, but that argument does not say

that important aspects of masterful performance cannot in principle be simulated by symbolic systems. To cite just one example noted earlier, Norman's (1982) distinction between goal-action and trigger-action schemata would seem to provide for the kind of situation-linked triggering that constitutes the "intuitive grasp" or "immediate know-how" of the Dreyfus expert. Norman even describes expert action in much the same terms as do the Dreyfus's. And both descriptions seem particularly appealing when complex perceptual and psychomotor expertise is the focus; whether physicists and mathematicians are best described this way is less clear.

Furthermore, some of the empirical results on experts cited by Glaser are consistent with the Dreyfus theory even though built from research based on symbolic processing assumptions; e.g., fast-access pattern recognition, multiple interpretation, automatic self-regulation and sensitivity to information in the situation, are characteristic of experts in both views. And, at least at the lower levels of the Dreyfus's scheme, it is not clear that instructional or assessment strategies would necessarily differ in the two views. The novice is taught to recognize and decontextualize key features, and to apply rules to these. The advanced beginner is helped increasingly to recognize meaningful elements in new situations and similarities with previous experience, and to apply rules to both context free and situational features. The competent performer is helped to use hierarchical plans, goals, and sequences of decisions based on the constellation of situational features that is meaningful in context. This progression is also implied by Glaser (1985). The novice needs to learn to distinguish the key surface features that connect to abstract rules, and to apply those rules in principled ways. Eventually, the more advanced learner can recognize the deeper structure signified by a constellation of situational features, and the appropriate

plans, goals, and decisions flow more or less automatically to principled solutions; i.e., the knowledge is proceduralized to the conditions of its use. These steps seem also consistent with the Dreyfus theory. Even the passage to proficiency, it would seem in both views, then involves much further experience with varying situations so that pattern recognition of what is needed in a situation becomes automatic, even if action still requires some rule-governed analysis.

The Dreyfus position does lead to strong admonitions against certain instructional and assessment uses of computers, particularly in the role of tutor or tutee. As a tool for adapting drill and practice, diagnosing bugs, or simulating phenomena that cannot otherwise be brought into the classroom, there are some cautions but no major problems in using computers. But the Dreyfus's believe that computer tutoring of learners is difficult, even in simple training domains; in school subject domains it is impossible. They say, in effect, that computers will never be made to understand what each learner already knows and can do, or what domain understanding really is, and will thus never be able to pinpoint misunderstandings or use learner background knowledge to reach breakthroughs. Worse, to make learners into tutors of the computer, as in LOGO and some other microworld systems, is to help progress from novice to advanced beginner while retarding later progress. This is so because analytic, verbalized, programmable knowledge is counterproductive to progress toward higher levels of expert performance. We need not detail this argument further here (interested readers should study Dreyfus & Dreyfus, 1986, pp. 122-157). The retardation hypothesis is after all an empirical question.

The Dreyfus's theory of expertise gives pause for concern about assessment design. But it does not follow that instruction or assessment needs to fall into the traps they see in

computerized tutoring. According to the Dreyfus scheme, one can certainly imagine instruction using computer coached drill and practice at some levels and human tutors at other levels, with computerized or other assessments at all levels. Assessing performance with simulation models need not require that the human learner behave like a computer. The argument about the psychology of higher levels, however, does suggest that assessment research should be targeted to those levels of expertise that are less controversial. Constructing useful diagnostic assessment in the range from novice performance to competence would in itself be a major accomplishment. Appropriately designed, it would help us study the levels of progression toward mastery, sharpen the psychological constructs involved, and perhaps eventually resolve or reduce the controversy.

There is a further point to consider. We believe that the emphasis on expertise has been overdone. Research on expert-novice comparisons has been useful in opening up new questions for cognitive psychology. But the definitions of "expert" and "novice" vary substantially from study to study, and all confound the many personal characteristics beyond knowledge organization that go into distinguishing those who become experts in a domain and those who do not, or have not as yet. For instructional and assessment purposes, it may be much more directly useful to build a psychology of the able successful learner as contrasted with the less able unsuccessful learner. Instruction and assessment should focus on how to promote progress from the latter category to the former, granting that it also should do nothing to retard progress toward expertise beyond the former. We may be able through this research to understand what makes for the development of highly competent high school or college mathematics or history students, for example, while never being able to pinpoint the idiosyncracies that make the mathematician or historian beyond

this.

Research on expertise, from whatever view, has not addressed this issue squarely. We call it the "aptitude question" for short. It is noteworthy that Herbert Simon, after a lecture at Stanford University in 1979, was asked, "Do you really mean that all it takes to become a chess expert is thousands of chess board configurations stored in memory?" He responded, "Well, of course, in the end it all depends on how much aptitude you have!"

Aptitude and adaptive instruction. Research on aptitude for learning from instruction has run in parallel with the growth of cognitive psychology since the 1960's. There have been points of contact, particularly in the process analyses of ability test performance previously mentioned. But the vision of integration of the two programs (Cronbach, 1957, 1975b; Cronbach & Snow, 1977; Snow, 1980, 1988) is still distant and vague. Instructional design proposals, whether of the DAS, LAS, or CAS sort include schemes for adapting instruction to individual differences among learners (see also Corno & Snow, 1986). LPS does also at another level. But the individual differences attended to are those reflected mainly in domain specific pretests or prior cognitive progress within the instructional program or domain of concern. Some attention has been paid to more general working memory capacity differences, notably by Case (1985). The Lesgold example shows how mathematical ability differences might be circumvented in science instruction. However, there are few other exceptions. On the other hand, there has been little progress in learning how to use aptitude profiles consisting only of a vaguely meaningful list of abstract scores. Assessment research and development now has the potential, at least, of learning how to guide adaptation of instruction using all the relevant aptitude information available in consort.

First, we need to be clear on the concept of aptitude, which is badly misconstrued when it is equated with terms such as "intelligence" or "scholastic ability". The "Scholastic Aptitude Test" is an aptitude test in the sense that it is designed to predict success in college learning activities; it is one measure of an important aspect of individual differences in readiness to profit from college-level instruction. Achievement in high school, particularly in certain courses, is another important aspect of such readiness. But achievement motivation is still another important aspect, and personal independence, flexibility, and responsibility are still other important aspects. In this instance, aptitude equals the complex of personal dispositions that constitutes readiness for the kinds of learning situations to be faced in college, or in a particular college; it does not equal what is measured by one or another test. In another kind of learning situation, readiness will be constituted somewhat differently; i.e., the aptitude complex will be somewhat different. As situations differ radically from college learning situations, or as they differ from any kind of formal "learning" situation, the aptitude complex will differ still further. But for any performance situation in which success can in principle be predicted, there will be an aptitude complex, i.e., a mix of personal predispositions that are propaedeutic, i.e., needed as preparation to achieve success in that particular situation (Snow, 1980, 1987, in press). In short, aptitude equals readiness for a particular learning situation; we might say it equals situated readiness, as a parallel to the concept of situated learning (Brown et al., 1989a; Greeno, 1988).

In the course of an instructional program, some sort of model of the momentary state of the learner needs to be built up. Next instructional steps can then be conditioned on that state so that each learner gets what is next needed. In other words, the student

model has a profile of learning progress in the goal structure of instruction to that point. Once it is agreed that other things known about the student's strengths and weaknesses should also be in the profile in addition to present state in the instructional program, the question is no longer whether to include aptitude information but rather what and how. To follow the Lesgold example again, prior development of mathematical ability is clearly relevant to learning in physics. For students strong in the relevant skills, physics instruction can proceed to use them. For students with some weaknesses in mathematics, instruction must either stop to remediate the weaknesses or teach the next steps in physics in a way that circumvents them. In other words, instruction acknowledges an aptitude-treatment interaction (ATI) and adapts as appropriate. But the adaptation is at a micro level rather than at the macroadaptive level of most previous research (Cronbach & Snow, 1977). Another example for physics comes from the research showing that fundamental misconceptions can arise when certain naive beliefs about the natural world are brought to instruction by different students (McCloskey, 1983). To the extent that these beliefs are difficult to change, they constitute sources of inaptitude that instruction must work on, or with. Presumably different treatments are needed for students with different belief systems.

Now suppose that we also know that the students to be treated differ in their concept of ability development (Dweck & Leggett, 1988); some are "mastery oriented", believing their ability will develop with learning and effort, whereas others are "performance-oriented", believing that since ability is fixed their aim should be to perform on learning tasks in ways that will exhibit ability to the teacher, or hide its absence. Should we not also seek to use this aptitude information in choosing next instructional steps? To the extent that this sort of aptitude difference exhibits differential influence on learning in

the program, obviously we should. The question is how? Assessment systems for instruction ought to be designed to detect important sources of aptitude and inaptitude and to suggest ways that instruction can capitalize on or circumvent them.

But how much aptitude information needs to be included? There are dozens of aptitude constructs that might be relevant to some instructional situation somewhere. And some kind of law of diminishing returns must operate; there will be a limit to the number of relevant, distinguishable, and measurable aptitude constructs for any situation and another limit to the kinds of instructional adaptations that can usefully be planned. One recommendation has been to concentrate on a priority list of major aptitude constructs for which much is already known in theory and in relation to learning from instruction (Snow, 1987, 1989a). Another recommendation is to concentrate assessment attention on inaptitudes--sources of resistance to effective learning progress--on the grounds that obstacles are more often critical for circumvention or remediation, whereas capitalization on strengths is less often critical. A third recommendation is to include those aptitude constructs that will be most useful in formative evaluation and redesign of instruction (Bunderson, 1969). All three perspectives converge on a minimum suggested list of aptitude constructs or core concepts including: fluid and crystallized ability, working memory capacity, achievement motivation and evaluation anxiety. For any specific instructional situation, additional constructs could be indicated.

Still a further important question concerns how such aptitude constructs can most usefully be represented as measures. The answer will differ for different constructs and for different kinds of instruction. For many purposes, a single score for fluid inferential reasoning ability may suffice; for some purposes, constituent analytical skills may be

needed. The need for multiple score profiles is probably greatest for crystallized ability and prior achievement, particularly with respect to reading comprehension and mathematics. And conventional standardized tests may not suffice for this purpose.

For example, cognitive theories of reading have been used to develop new measures of reading competence that promise greater diagnostic and instructional utility than existing tests of reading (Calfee, 1977, 1982; Carr, 1981; Curtis & Glaser, 1983; J. Frederiksen, 1982; Johnson, 1983; Schwartz, 1984). Calfee's (1977, 1982) work serves as one example of such efforts. In contrast to the traditional view, Calfee has suggested that reading assessment can be more accurate and more informative if it is based on the idea that reading ability comprises of several independent rather than highly interrelated cognitive processes. He maintains that the high degree of interrelatedness of subtests in most reading assessment batteries results from "a lack of systematic design and contamination by uncontrolled factors that in the aggregate cause most tests to converge..." (Calfee, 1982, p. 166).

Calfee's assessment system evaluates student performance in seven skill areas: word attack, word recognition, word meaning, sentence fluency, sentence flexibility, reading comprehension, and listening comprehension. Based on these results, profiles over the seven different areas can be computed that reflect the pattern for the group or the individual under investigation. Research indicates that the use of these profiles can provide important insights to teachers regarding skill differences between reading ability groups (Calfee, 1981; Calfee & Spector, 1981). In particular, when students classified either as learning-disabled, or bottom quartile readers but not disabled, or average readers were assessed in this manner, specific patterns of weaknesses and strengths related to ability

level were found to exist in the various skill areas. The value of such information for instructional design in reading is clear. Rather than simply repeating lessons or slowing the rate of instruction, remediation can be given to lower ability groups in specific skill areas. The profile approach can also have important implications for the individual student in other instructional domains. For example, even when two students are both identified as average readers, their individual profiles in the various skill areas could be quite different. Again, this might indicate the need for two very different forms of instruction to circumvent the different weaknesses.

Achievement motivation and anxiety are also constructs that have been subdivided by modern theoretical developments. The Dweck-Leggett (1988) distinction between mastery versus performance motivation is one example given earlier. Entwistle's (1987; see also Marton et al., 1984) work results in a related distinction, between approaches to learning characterized by deep versus surface versus strategic goals. Similarly, mathematics and computer anxiety have been distinguished from evaluation anxiety, and the latter subdivides into worry and emotionality components.

What further constructs and what level of abstraction will suffice for each is a matter of hypothesis and evaluation in each instructional domain. In one, weakness in visualization may be a crucial addition. In another, learner's choice of instructional material based on interest preferences may be worth investigating. In any event, much recent research argues that aptitude information should not be ignored in instructional design or in learning progress assessment.

No one program of research and development on SIIA can be expected to encompass the whole network of aptitude, learning, development, and achievement

constructs represented by Snow (1989c), much less the elaborations and additions that are possible. Nor should one try. But no research and development agenda should proceed without recognition of the rich psychological network within which it works. Some SIIA may be made adaptive to some of these psychological variations, if they are in attention as work proceeds. And evaluations of SIIA, if not their development, need to represent all relevant constructs.

Diagnostic Assessment for Instructional Use

Diagnosis. The ultimate aim of SIIA is to provide diagnostic assessment for use in understanding the present state of learner progress, communicating it, and making instructional decisions about next steps. This sort of assessment differs from the traditional model of educational measurement in three fundamental ways.

First, diagnosis serves instructional placement or classification decisions rather than student selection decisions, in the sense defined by Cronbach and Gleser (1965). That is, it guides the assignment of a student to one of two or more alternative instructional treatments aimed at common endgoals rather than to accept-reject (or grading) categories. In Figure 1 as previously noted, the inset indicates six alternative next steps based on the diagnosis provided by reference task B4. These are six alternative treatments or routes to a common endgoal. The choice among them is a classification decision, aimed at adapting the next instructional step to each student's present state.

Second, the classification decision is made at what has been called a microadaptive rather than a macroadaptive level (Cronbach & Snow, 1977). Whereas most research on ATI has aimed at classification of students into alternative instructional treatments on a month-to-month scale, microadaptive assessment must support classification decisions on

a week-to-week, day-to-day, or perhaps even an hour-to-hour scale. In the extreme, adaptive instruction is made "response-sensitive" as in CAS; the decision rules for choosing alternative next steps control the frame-by-frame sequence on a scale of seconds and minutes (see Atkinson, 1972; Vanderlinden, 1984; Wenger, 1987).

Third, the classification is based on an interpretive theory. It is an attempt to explain why a student performed in one way rather than another, and should thus be treated in the prescribed way. In other words, the term "diagnosis" means classification into an explanatory category as well as into a particular treatment in instructional psychology, just as it does in clinical psychology or medicine.

There would appear to be a heavy burden of validation to bear in justifying decision rules of this sort. Ideally, each instructional choice would be proven optimal for students with particular patterns of response on each reference task. At the present state of the field, however, we do not know how to do this at the finer grains of instructional adaptation. And perhaps validation at this microadaptive level is not really necessary; there may be approximations at some higher level of aggregation. For example, Atkinson (1972; see also Chant & Atkinson, 1973) sought validation of an instructional optimization program by showing that the aggregated effect on achievement of its decision rules was superior to that obtained by learners making their own next-step decisions, and to that obtained by fixed sequence control groups. Suppes (personal communication, 1985) conducted evaluations aimed at the level of student trajectories over the network of tasks in a course, rather than at individual tasks. Snow et al. (1979; see also Snow, 1980) showed how learning curves within a CAS unit could be combined with indices of learner activities during the adaptive sequence, and aptitude scores gathered before it, to suggest

improvements in the pattern of decision rules for particular students. Bunderson (1969) demonstrated how detailed program changes could be guided, over several program revisions, to eliminate molar ATI effects.

Variations on these techniques may be helpful in formative evaluations of SIIA. But the question of diagnostic validity will remain troublesome. General discussions of validity and of diagnostic testing have only begun to address the problem of diagnosis for instructional decision-making (Messick, 1989; Nitko, 1989). It is also clear that validation of diagnostic assessment must link constructs and evidence with value considerations involving social and personal consequences (Cronbach, 1988a, 1988b; Messick, 1989).

Types of diagnosis. The problem may be analyzed, and perhaps ultimately simplified, if we recognize that next-step decisions at any level of adaptation depend on three types of diagnosis. Further, each type can be expressed in the form of one or more condition-action statements and a priority order can usually be imposed on the resulting list of statements. As previously noted, some decision alternatives require tight confidence bands, considering the consequences of an incorrect decision. For others, the choice between alternative next steps may not be so consequential. For simplicity, we refer to the three types as readiness diagnosis, remedial diagnosis, and resistance diagnosis, but other terms would serve as well.

Return again to reference task B4, and suppose that we have adopted the model of learning proposed by Norman (1982), at least for this task and the prior instruction leading to it. The task has thus been designed to provide scoring indicators associated with Norman's three phases of prior learning--accretion, structuring, and tuning. As suggested in Figure 2, the key distinction for readiness diagnosis is whether accretion,

structuring, and tuning phases are complete or incomplete. The key for remedial diagnosis is whether B4 performance indicates that accretion or structuring is faulty. Resistance diagnosis provides subsidiary information to help examine generally poor performance on B4, or to help decide which next step to take when B4 learning is complete. Now consider each type of diagnosis in more detail.

Insert Figure 2 about here

Readiness diagnosis first concerns whether or not a student is ready to move beyond this reference task. Measures are thus required to identify various forms of incomplete learning. Incompleteness might be indicated by errors of omission, guessing, or "don't know" responses, or by halting, slow production of an explanation or performance of a task component requiring procedural skill. If learning is judged to be significantly incomplete, then some combination of review, coaching, and practice is needed. Accretion might be relatively simple to complete through additional instruction and review of missing facts, concepts, or procedures. Incomplete structuring would require further instruction, or carefully coached practice, to ensure that appropriate as well as complete structure emerges. Tuning probably can be completed with more extended practice. If learning is initially complete, or after it becomes complete with this additional instructional attention, then readiness diagnosis identifies the optimum next step for instruction. The choice will be the step that capitalizes on each student's strengths and, if necessary, that circumvents any known general weaknesses that can not readily be removed at this point. Resistance information is used to aid this decision. If the step to B5, for example, requires a level or

kind of mathematical or spatial ability not presently possessed by a student, the step to A5 might bypass this source of resistance to learning progress, at least temporarily. Alternatively, a step to A4 might be aimed at building the needed skills before B5 is attempted.

An example of this sort of diagnosis in CAS was given earlier. That program uses information on the present state of a student model, in relation to the curriculum goal structure, but also aptitude information--the student's mathematical ability--in choosing a next step that avoids a source of resistance to learning. Alternatively, as noted, the program might shift to a line that builds mathematical skill directly. Readiness and resistance diagnosis have to be used in consort, because instruction cannot work on the whole student model at once; it must sidestep some weaknesses while working to remove others.

Remedial diagnosis distinguishes faulty accretion or structuring from learning that is merely incomplete. The form and severity of the problem would be indicated by various kinds of errors of commission. Thus, the reference task must provide means of identifying the major known sources of error, and should be arranged to detect important but idiosyncratic errors as well.

Decisions as to what remedial instruction is needed, or to what previous instructional steps a student should return, depend crucially on diagnosing the sources of error, or faults, in present performance. Faulty structuring is likely to be the highest priority for assessment research because such deviations may be the most damaging to later progress and the most difficult to remediate. A student who has somehow developed a major misconception or badly flawed procedure, or who has built new structures onto faulty prior beliefs or

procedures, may be troubled not only in B4 performance but also in many later instructional steps. Faulty structures may even be kept "underground" that remain undetected in a student who otherwise passes conventional unit tests, may appear much later as deep and limiting misunderstandings (Marton, 1983; Norman, 1982). Remediation of accretion errors is of course also important, though presumably errors of fact or simple concept, or of simple procedure, are more readily repaired. An implication of this analysis is that reference tasks should be designed to provoke errors likely to indicate important faults if they are present.

Resistance diagnosis, as noted above, provides subsidiary information in support of both readiness and remedial decisions. It derives from records of prior instructional performance, ability test profiles, and perhaps background interviews, questionnaires, and other sources, and is designed mainly to identify more general weaknesses and resistances relevant to planned instruction. Low present ability in an area not explicitly taught but required for some instructional goals was one example given above; this observation might aid in readiness decisions about alternative next steps following completion of B4. Ability profile information might also help elaborate remedial diagnosis and the choice of next steps in this direction. Conative as well as cognitive resistance information is also useful in these ways. For example, indicators of achievement motivation, anxiety, unrealistically low self concept, or the belief that ability cannot be improved might help explain a particular student's performance in B4. They also might suggest a shift, following B4, to a parallel step A4 designed to build further self-confidence in improvement on tasks of equal rather than increasing difficulty (as in A5 or B5). Certain kinds of learning style and interest measures might also be useful in this way.

Resistance diagnosis may be of secondary importance in most cases, compared to remedial and readiness diagnosis. But SIIA research and development should include at least the most obvious indicators of such sources of resistance. Mastery in an instructional domain is a psychological construct with both cognitive and conative constituents, and both broad and narrow implications for performance. It follows that the promotion and evaluation of learning progress toward mastery needs all three types of diagnosis.

Levels and types of faulty learning. A further implication is that a crucial first step in SIIA development in any domain is to build a catalogue of errors typically observed in performance at various levels of progress. Reference tasks should be designed with alternative possible sources of errors in mind. The enumeration of likely errors should include error expectations from the instructional experience of teachers and experts, as well as studies of reference task performance. To the extent that reference tasks are built from cognitive task analyses and simulation models, it may also be possible to include routines that reconstruct or even generate possible errors in the tasks, to elaborate hypotheses about the sources of certain kinds of errors.

The experiences and methods of research on CAS may be especially useful here. But there are some serious limitations. In CAS research, simulations used to guide diagnosis are mainly of intermediate levels of performance. The performance of extreme novices and experts is not well simulated. Also there are various theories of errors, but each tends to be limited to a particular tutoring system for a particular domain. For example, using the ACT* theory of skill acquisition, Anderson and his colleagues have designed tutors for LISP programming, high school geometry, and algebra (see Anderson et al., 1985; Anderson & Reiser, 1984; Boyle & Anderson, 1984; Lewis, 1989; Reiser,

Anderson, & Farrell, 1985). The systems are based on a model-tracing methodology. That is, at each point in the problem solving process, the system uses its database of rules to generate all possible next steps, both correct and incorrect, that the student might take. Whichever rule corresponds to the student's next action is interpreted as reflecting the student's reasoning; if it is incorrect it is immediately corrected. When the tutor is unable to account for an action, the system signals its lack of understanding. If, after several attempts, no model can be created, the tutor simply gives up and tells the student the appropriate next step to take. Thus, it may fail to diagnose some potentially important aspects of novice performance.

The model-tracing methodology has other questionable features (Wenger, 1987). It appears to be an effective method for introductory level instruction in highly structured domains because the highly directive nature of the system helps keep the inexperienced student moving forward toward an appropriate solution, but it provides correction without diagnosis outside of its own rule system. Moreover, the restrictive nature of the system seems to provide no way for beginning students to develop self-regulatory experience in the domain. And this aspect of the tutor also makes it especially undesirable for the more advanced student. Students at this level may need to be given the freedom to make mistakes and explore solutions that often result in increased learning and understanding.

Wenger's discussion of this limitation relates back to our discussion of alternative views of expertise. Tutoring systems based on skill acquisition theory will not necessarily lead to the development of expertise in the subject matter because the acquisition of expertise and the acquisition of skill are not equivalent. "Whereas skill acquisition can be tested by straightforward performance measures, expertise is a much more subtle notion:

it implies an in-depth understanding of the domain that includes, but goes beyond mere performance" (Wenger, 1987, p. 302). Thus intelligent tutoring systems, or SIIA, that have the goal of developing in-depth understanding must provide opportunities for learning and exploration across a variety of situations, not just error correction within a narrow structure.

Other systems seem to have similar limitations. The methods of diagnosing errors used by DEBUGGY and PIXIE noted earlier are also limited to relatively simple problems, in a narrow domain, so that relatively complete bug libraries can be created. In broader, more complex domains complete bug catalogs may be impossible to construct. Yet student bugs become more critical as problems become more complex, and here a list of single bugs, even when used in combination, is likely to be of little use. Work by Bricken (1987) and others (Neches, 1982) has shown that learners make a wide variety of "sloppy" or idiosyncratic errors, some of which they themselves are able to recognize and correct. This suggests that many errors are not due to either missing or incorrect rules.

In short, there are not likely to be general theories of faulty learning, or of errors at different levels of learning progress. The development of reference tasks in a particular domain will have to build its own theory about the kinds of faults possible and what they mean for subsequent instructional steps. There are likely to be not only different sources of error at different levels, but also mixtures of domain-specific and general sources of error. At this point, we can only demonstrate that there are many kinds of faults possible in learning and cognitive performance in a domain, and emphasize that the identification and classification of these faults is the foundatic of any attempt at diagnostic assessment.

Table 3 lists some of the kinds of faults to be expected. The list is neither exhaustive nor mutually exclusive. It merely exemplifies the complex problem to be faced

in reference task design and in diagnosis. It is based on logical considerations plus a review of discussions in Mandl and Lesgold (1988), Psotka et al. (1988), and Wenger (1987). It will be important also in reference task design to recognize behavior that, while not faulty, can indicate weaknesses or inflexibilities in learning. A symptom list should be generated, to include signs of weakness such as verbatim recall, small chunk size, and algorithmic repetition in student performance, as well as symptoms of conative and affective states that suggest nonoptimal learning.

Insert Table 3 about here

There is a further point regarding levels of diagnosis that bears emphasis because it has received less attention in research to date and yet may be critical, especially in the diagnosis and remediation of faulty structuring. CAS research seems to focus on individual errors, or similar errors in series. But some deeper structuring problems, such as those involved in fundamental misinterpretations or conflicting personal beliefs, are probably to be detected in patterns of performance, not in individual errors. This requires aggregation of evidence to depict diagnostic constructs, useful not only at the level of choosing next instructional steps from a particular reference task, but also at the level of planning and communicating about progress in a course of instruction.

Psychometric Problems

Before addressing the new psychometrics that will be needed in work on SIIA, three other issues need to be faced and removed, for they will otherwise plague SIIA research at all levels. One is the problem of the measurement of change, with its attendant

psychometric difficulties. The other two concern the methods used to obtain so called "dynamic" or "authentic" assessments.

Measurement of change. Whenever the terms learning, development, progress, or gain are used, and discussion centers on assessment, it often seems to follow automatically that the psychometric problem should be viewed as one of measuring change, i.e., quantitative difference across time. After all, learning and development are defined in terms of change. But this view derives from the assumption that there is an underlying, parametric continuum that is common to each time-point or stage of learning; the learner progresses or gains along a single quantitative scale. Even when learning is recognized as multivariate, each latent parameter at one point in the learning process is assumed to have its isomorph at another. The conception is that "the same variable" exists at two or more points in time for every learner. Statements can thus be made about learning rate or relative learning gain for different learners or instructional treatments. The psychometric troubles then arise from questions about the reliability, validity, and scale of measurement of change scores computed for individuals as differences across time on this continuum. There are usually also related questions about floor and ceiling effects. It is not clear that these questions can be answered meaningfully for such scores, at least in most situations.

The problem comes from the test design principles and mathematical models for measurement conventionally applied, and can be removed if these conventions are set aside in favor of a different view. Latent-trait and item response theories may apply well to some individual assessments at single points in time. They may even apply to some kinds of change measurement; change in one or more skills or capacities each univocally measured on a clearly unbounded interval scale might be the best example. Advances in

psychometric research now offer ways to address change measurement in such situations, using specialized latent trait models (Embretson, 1987) and growth curve models (Rogosa, Brandt, & Zimkowski, 1982). But we do not believe that the assessment of individual learning progress in instruction is a situation likely to be well fit by such models.

Complex learning of the sort addressed here is better viewed as a multivariate progression across qualitative states. Each of these states can be described using both quantitative and qualitative methods, hopefully in rich enough detail to guide interpretations and decisions about individual learning progress. But only in superficial ways do "the same variables" apply to each state or stage of this progress. All of our previous discussion of the nature of cognitive and conative school learning supports this view. Thus, we think the position recommended by Cronbach and Furby (1970) still applies. Change due to learning from instruction is not to be directly assessed. But significant states of progress can be separately assessed in ways that afford rich description as well as prediction from one to another (see also Cronbach, 1975a, 1982b, 1986). The problem of measuring change then does not arise.

If this view is adopted, then it follows that research on DAS and LAS ought to avoid gain scores. Each system offers enriched descriptions of learning progress without them, and evaluation of these systems may be misguided by them. CAS and LPS development do not so far depend on gain scores; they should not be used in further work.

Dynamic assessment. We also oppose the contrast of "static" with "dynamic" assessment. As noted earlier, various LAS developments have been described as seeking "dynamic assessment" (Lidz, 1987). Embretson (1987, p. 143) defines this venture as "... attempts to modify the performance level of an examinee by the design of the testing

materials or test administration procedures." Bunderson et al. (1989, p. 369), on the other hand, follow an analogy to physical measurement to define the contrast more generally: "If static measurement is the specification of a point, or points, in an educationally relevant measurement space, dynamic measurement is the specification of a trajectory, or path of points, over time. If a point defines a position along a relevant scale, then a trajectory defines changes in position over time."

Such definitions have problems. They are fine if "dynamic testing" simply refers to materials and administration designs that promote learner change, and "dynamic measurement" is a path of static points that describes learner change in psychometric space across time. No individual change computations need be implied. But these definitions do not agree: a path of static points based on conventional tests would seem not to be "dynamic" for Embretson and other LAS researchers, whereas a static score based on performance in a Campione-Brown hint hierarchy would presumably not be "dynamic" for Bunderson and colleagues. Furthermore, in physics "dynamic" implies the possession of a differential calculus, which psychometric theory certainly does not now possess. In psychology, "static" implies that the learner's performance is not dynamic; one has only to observe conventional administrations of individual intelligence tests or examinee eye movements while producing responses to conventional ability or achievement test items to recognize that "static-test" performance is in no sense static (see e.g., Snow, 1978, 1980). The numerals in a resulting test score or the words in a clinician's report are static only in a trivial sense. We thus believe the terminological contrast between static and dynamic should be abolished. At the least, there needs to be much further clarification of the conceptual distinction as it applies to diagnostic progress assessment, and the degree to

which direct change measurement is actually part of the concept.

"Authentic" assessments and portfolios. Another development to be noted is termed "authentic" assessment because it uses learning tasks that are recognized as being legitimate, representative tasks of a subject discipline and that serve assessment purposes as well as learning goals. Tasks are chosen to require and exhibit particular kinds of ability developments and to serve as reference points for teacher observation and teacher-student communication about progress. For example, Gardner and Hatch (1989) choose tasks emphasizing one or another of Gardner's (1983) seven hyphothesized kinds of intelligence. Gitomer (1989) chooses tasks to reflect students' development of performance in the arts. Student productions in either case can be collected into portfolios for review and criticism by the student, teacher, or peers at various stages of progress.

A further example of the type of task used in portfolio assessment is drawn from the imaginative writing component of the Arts PROPEL Project. Howard (1990) describes a task called the first written reflection in which students are asked to "reflect upon themselves as writers". The task requires students to identify and discuss one thing in their writing that they believe they do well. They then must describe one thing that needs to be improved. Such analysis necessitates that students not only produce a concrete piece of writing but also engage in valuable self assessment. Subsequent written reflection tasks build in this dual-purpose objective; students reflect on the strengths and weaknesses of their work through a process of self-criticism while continuing to produce additional writing samples. The tasks becomes increasingly elaborate and demanding, culminating in a critique and comparison among work samples in the students' portfolios.

These approaches may provide a useful base for instruction and assessment, but

the tasks are not designed to be diagnostic following the definition used here. There is no objective scoring or classification system built into the tasks, and no explicit decision points for instructional steps. Rather, diagnostic interpretation and responsibility for instructional adaptations rests solely with the teachers and students. Research in this line faces the same difficulties that have plagued previous work on clinical decision-making in contrast to statistical predictions.

New psychometric models. The above discussion should not be taken to discourage research on new psychometric models useful in learning assessment. There are important new developments along several lines in this direction (see, e.g., Frederiksen et al., 1990, in press). Some approaches use item response theory in combination with other mathematical models (see, e.g., Embretson, 1990; Tatsuoka, 1990). Some explore directly the integration of cognitive and psychometric models in the context of literacy tasks such as those used in LPS (Sheehan & Mislevy, 1989). Some even provide a specialized kind of change measurement using homogeneous groups of items arranged to depict progress through a knowledge hierarchy (Rock & Pollack-Ohls, 1987). But we wish especially to encourage a more divergent search, at this stage, for models that describe qualitative structures of knowledge and skill, related conative states, and shifts among them, without the need for quantitative change measures.

One example in this direction is seen in the work of Kyllonen et al. (1984), Mislevy and Verhelst (1990) and Ohlsson (1984a) on strategy differences and strategy shifting during learning across a series of problems. Another example is the modeling of skill pattern progressions across instruction studied by Wiley and Haertel (in press). The patterns show that skill acquisition can proceed through qualitatively different routes as

well as sequences, and suggest that persons and instructional treatments can differ dramatically in them.

Progress assessment through a domain topography. We stress this last point because assessment models are needed for diagnosis, beyond the level of individual reference tasks, at the level of progress across tasks, units, and courses. This brings up the grain-size problem again. We earlier noted that fine-grain analysis of individual task performance might well not aggregate appropriately for use in teacher interpretations or discussions at weekly or monthly levels. We also discussed the need for confidence bands connected to instructional next-step decisions at the level of reference tasks; these sorts of probability statements also are unlikely to aggregate meaningfully. A different level of description seems to be needed to characterize learning progress across a domain topography. And different domain topographies may require quite different assessment models at this level, as well as at fine-grain levels.

At molar levels, it would seem that interpretive constructs are needed that reflect qualitative shifts in conception, strategy, viewpoint, or attitude in a domain, as well as pattern or profile shifts in skill acquisition. These shifts cannot be described simply as accumulations in numbers of concepts or strategies or skills possessed. They rather should describe progressive steps or stages expected during normal development in a domain and important kinds of deviations therefrom. Such interpretive constructs need to be cast in substantive, domain-specific as well as psychological terms.

We cannot attempt such definitions here, because that is a matter for theoretical consideration and construct validational research in each domain. We expect, furthermore, that no abstract definition of learning progress constructs are likely to hold across domains,

and even end-goal constructs are likely to differ in the abstract. For educational purposes we might agree, for example, that mastery in mathematics means thinking like a mathematician, which in turn means embodying the fundamental structures and processes that characterize understanding and reasoning in mathematics and using this knowledge profitably in other school and life pursuits. In civics, world history, art, or physical education, however, we are not likely to agree that mastery should mean thinking like a politician, a world historian, an artist, or an athlete. Each domain will have its own definitions of progress constructs worked out in the choice of instructional psychological goals for a segment of a population to be served. The goal lattice structure and the network of reference tasks, or learning progress map, for each domain provides an operational definition of progress levels in that domain.

One might try to equate levels of learning progress with levels of expertise. On reflection, however, we think the novice-expert continuum is potentially troublesome for interpretive constructs in most of the domains where SIIA is likely to be developed. The troubles come from both substantive and terminological concerns. At the expert level, for example, thinking like a mathematician or historian does not mean becoming a mathematician or historian. Yet the constructs used to interpret expert performance indeed do come from research describing mathematicians and historians, in contrast to various categories of nonexperts. Such research confounds the cognitive contrasts of interest with all the other personal characteristics correlated with it and glosses over the qualitative distinctions among the groups studied within and across domains. There is also the threat that abstractions such as "expert" and "novice" become social labels for people. At lower levels, if an abstract description of levels of progress is needed for ease of

communication, then we think there are serviceable terms with everyday meanings. "Beginner" and "advanced beginner" have such meanings and do not carry the demeaning tone of "novice". "Competence" and "advanced competence" are also everyday terms; though they do not have absolute meanings, it seems possible that acceptable psychological definitions can be worked out in the abstract, in keeping with the evidence and hypotheses advanced by Glaser (1985), Dreyfus and Dreyfus (1986), and others. Also, for practical purposes at least in high school academic courses, describing a level as "advanced competence in mathematical thinking" carries an implied relation to "advanced placement" or "honors" courses. Though this does not solve the problem of defining psychological goals and constructs, it does at least convey a public sense of level.

For purposes of research communication, at least, we can use abstract levels to sketch what learning progress maps and some diagnostic constructs might be like. Again we rely on the topographic metaphor. Levels of progress are likened to levels of elevation marked on a fairly coarse map, and a given learner may have a profile that shows different elevations in different regions of the map at any point in time. The learner also has a route over time which, in keeping with our diagnostic terminology, we think of as the path of least resistance. Different learners will thus take different routes through the map, and will have different elevation profiles at different points in time. For simplicity here, we assume that every learner eventually performs on every reference task, though each may take a different route through them. This need not be assumed, however; a prediction system in the task network might estimate that learners showing marked success up one region of the mountainside need not return to fill in another region below and to the east, for example.

Figure 3a provides a schematic three-dimensional view. There is a horizontal map at the base consisting of a goal lattice layer and a corresponding network of reference tasks. Thus, the horizontal dimensions are the array of end goals and viewpoints, and the progression of subgoals and reference tasks for each, as worked out using whatever educational wisdom and previous research exists in the domain. Again, for simplicity here, the reference tasks are tiered into those suitable for beginner, advanced beginner, competence, and advanced competence levels. The vertical dimension also shows these four rough elevation levels of learning progress. Position A1 is one reference task in the network; we can think of it as a small nest of arithmetic word problems of a certain type, designed as in LAS for example. Imagine that it affords assessment of levels of skill in four aspects of problem solving as depicted in the Heller-Greeno example from Table 2, as well as the learning propensity and transfer measures of LAS. Thus, it provides elevation estimates for several aspects of this task. Position A2 is a reference task at the next level of progress, similarly designed. The performance indicators for both tasks show complete success for this student on this route. The instructional decision at this point would be whether the learner should continue on this route or return to beginner levels in regions B, C or D.

Figure 3b shows the same cube, but with one learner's complete and perfect progress through the course shaded in. An average of all learner's surfaces would look the same in a perfect course. In Figure 3c, however, variations in class average elevation would identify regions of effective instruction and easy reference tasks (ridges and peaks) and also regions of ineffective instruction and difficult reference tasks (cliffs and gullies); these are targets of difficulty deserving further research and development. The horizontal

plane atop the cube depicts the expected performance of mathematicians, if this domain is mathematics. The horizontal floor of the cube depicts the expected performance of persons who are not ready for this course.

Figure 3d suggests the profile at some point in time of an individual learner who, through anxiousness or lack of confidence perhaps, prefers to master all tasks at beginner levels before moving to tasks at more advanced levels. Figure 3e suggests the profile of a learner at the same point in time who prefers to reach deep understanding and skill in one region, skirting other regions at first because of real or perceived weaknesses there. In an actual instructional domain, we could add substantive description to make these into interpretive constructs. Note, however, that these constructs, as well as those describing average course performance in Figure 3, would offer interpretations at a molar level of grain. They are not simple aggregations of scores from reference tasks, and they leave out information concerning individual routes through or repetitions of reference tasks, or confidence bands for individual steps. The average surfaces could include indications of variance in the class at each step, but we have omitted these here.

Insert Figure 3 a, b, c, d, e about here

Notes on conative diagnosis. The problem of conative assessment in instruction has received nowhere near the research attention it deserves, so SIIA development must start with conventional techniques here. Various motivational, volitional, and personality constructs can be assessed by questionnaire, projective, or observational methods. Some of these might be redesigned for systematic use by teachers, and of course periodic

questions, self-ratings, and the like might be incorporated into instruction or reference tasks. However, new research needs to pursue methods that can be linked more directly to student performance (see Kanfer, Ackerman, & Cudeck, 1989). This last section singles out two such projects as examples.

One line of work aims at assessing central constructs in Kuhl's theory of action control, namely, self-regulatory efficiency and metamotivational knowledge (Kuhl & Kraska, 1989). For self-regulation assessment, a computerized, divided attention technique is used wherein students perform a main, symbol recognition task and are rewarded for speedy accuracy, and thus concentration. The screen also contains a simultaneous distractor episode which influences students rewards but over which they exercise no control. Though students readily form a commitment to concentrate on the main task, maintaining this commitment is difficult in the face of the influential distractor. Self-regulatory efficiency is then indicated by a measure of variance in interresponse times, reflecting increases in speed of performance to offset the slowing that accompanies glances at the distractor. This measure avoids the problem of confounding action control efficiency with motivational change toward disengagement, and shows promising correlations with teacher ratings of attentiveness, concentration, and independence. The measure of metamotivational knowledge is a test consisting of classroom scenes with captions representing different control strategies for maintaining the intention-action commitment to learn. Students respond by indicating which pictured person cannot maintain the commitment to learn. The measure shows developmental trends consistent with theory and also yields promising correlations with teacher ratings; the knowledge test and the performance efficiency measure also correlate. Kuhl's research is now advancing to the development of similar

assessments of older age students and instructional interventions designed to improve self-regulatory strategies. Teacher use of such instruments, combined with their own observations, can easily be imagined.

Another line of work is suggested by the efforts of a team of Belgian researchers to develop a free-response personality inventory (Claeys, DeBoeck, Van den Bosch, Biesmans, & Bohrer; see summary by Snow, 1989b). The approach uses a computerized dictionary of self-descriptive adjectives, with a system of weights reflecting expert scaling of each adjective on each of several personality dimensions; respondents are assigned score profiles based on the weights of adjectives they use to describe themselves. Validity studies suggest that free-response conditions may activate personal knowledge structures in ways that conventional questionnaires used alone do not, and thus produce more valid response. Free response allows individuality of self-description and perhaps a more intensive conscious search of personal knowledge. It is notable that the personality dimensions so far studied in this way include the learning-related constructs of conscientiousness, achievement motivation, anxiety, and self-confidence. One can imagine descriptor systems designed along these same lines but focused more directly on instructional motivations, interests, and self-perceptions as well as on particular learning activities. It might even be possible to collect such scaled self-assessments periodically as instruction proceeds, to produce a richer and more integrated description of cognitive and conative aspects of personal knowledge growth.

Toward Domain Topographies For Instructional Assessment

This final section gives a summary agenda and recommendations for SIIA research and development, and then adds a concluding overview. It is part summary and part

further example of steps we now think work on SIIA is ready to take, and considerations that need to be in mind as they are taken. For ease of reference, it is organized as a numbered list, categorized within the major phases of a projected research and development program. The list is intended to suggest a sequence that could be followed, though it emphasizes the earlier steps toward establishing domain topographies and networks of reference tasks, rather than the many engineering design problems that must ultimately also be faced. At this stage, of course, even the early order is approximate and incomplete.

We see distinct advantages in bringing work related to our four example systems together. As noted earlier, the four are not competitors. The approaches provided by DAS and LAS represent intensive integrations of assessment and instruction at the level of reference tasks or families of such tasks. CAS provides this as well, but also moves toward encompassing larger chunks of curriculum. LPS seems to provide an overarching framework at the course level into which the other work might fit. Accordingly, we have framed the proposed research and development agenda on the assumption that this overall direction is potentially the most fruitful choice.

A Summary of Agenda and Recommendations

1. Choosing domains and end goals.

1.1. Instructional domains that have been more fully studied to date will afford a firm foundation and a faster start toward the early development of SIIA. The mass of prior cognitive psychological analysis of mathematics tasks makes this domain a first choice. The topics offering the deepest base of prior work are arithmetic computation, arithmetic word problems, algebra, and geometry. There are existing DAS, LAS, and CAS examples in all these areas. Mathematics is a good first choice because it is a major target of

difficulty for teachers as well as learners, and because mathematicians and mathematics educators have spent considerable effort in recent years debating and clarifying different conceptions of its critical end goals.

1.2. Concentration only on mathematics would be severely limiting in the long run, however. The special character of mathematics learning and instruction will produce features and principles for SIIA development that will not translate well to work in other domains. While this specialization problem may be inevitable in any domain, we think it a particular concern in mathematics. We recommend, therefore, that at least two and preferably three other domains with different structural characteristics should be addressed in parallel with mathematics. Candidate domains would seem to be science, history, and literacy.

1.3. Selected topics in science have also received relatively more attention in DAS and CAS research, and may be the best second choice. Here the most studied topics are electricity, Newtonian mechanics, and several elementary science concepts involving time, motion, and distance problems, balance beams problems, and biological classification. In addition, there is a range of new research on the assessment of science achievement (e.g., Gong, 1988; Martinez & Lahart, 1990; Shavelson, Pine, Goldman, Baxter, & Hine, 1989). The choice of science as a second domain has the additional advantage of allowing exploration of interstitial instruction and assessment; learning mathematics and science together should help make each domain more meaningful.

1.4. The third domain could be either history or literacy, or the two taken together. Only a few first steps have been taken to build a cognitive psychology of history (Wilson, 1988; Wineberg, 1989, 1990). It is another obvious target of difficulty, however,

with a purpose and structure different enough from both mathematics and science to be illuminating for SIIA. Another valuable choice is literacy, both elementary and adult, again because there is important prior work (Kirsch, 1987) and important teaching and learning difficulties. There are also interstitial possibilities among these domains; the two can complement one another.

1.5. Whatever the domain a next crucial step is goal analysis. We have given an example from Lesgold's (1988) work of a goal lattice that incorporates multiple end goals and multiple teacher and expert viewpoints about them. We believe that reliance on teachers and domain experts in developing goal and viewpoint descriptions is indispensable. But the purpose is not only to obtain informed opinion. It is to examine teacher and expert explanation and understanding itself as two kinds of expert performance. The Case-Bereiter task analysis of expert and successful student performance is one important approach to this step. The aim is to obtain statements of the psychological goals of instruction and viewpoints about them.

1.6. The consideration of goals should not be allowed to converge only on proximal, cognitive objectives explicitly statable in advance. There are usually longer-range, higher, emergent, and divergent goals that need to be described even though they may be difficult or impossible to specify in measurable terms at the start. Some concern deep understanding of the central conceptual and procedural structures in a domain. Some involve conative aspects expressed in terms of motivation, volition, interest, and attitude. Some concern transfer relations, both positive and negative, between adjacent domains. In the view of some teachers, domain experts, and researchers, limiting SIIA development to the goals that are most readily measured would be exactly the wrong way to start.

2. Identifying intermediate goals and needed reference tasks.

2.1. For each domain, a psychological topography is needed. This can be built initially using descriptions of instructional goals, theories about curriculum structure, expert viewpoints and experience concerning what concepts or procedures are particularly difficult to learn or to teach, arrays of existing achievement test items with known difficulties, and tasks or problems for which cognitive analyses or simulation models are in hand. Through continuing cognitive analysis across these arrays, the aim is to produce a topography of intermediate psychological goals for instruction. Reference tasks are then designed to fit the key points in this topography, i.e., to form a learning progress map as envisioned in the work on LPS.

2.2. General taxonomies of instructional objectives and the content-by-process tables typically used for achievement test specifications are of doubtful use in building such a topography. They are likely to be counterproductive as starting points in this regard, unless accompanied by substantial construct validational research.

2.3. Although specifications of behavioral objectives or of prerequisite intermediate steps or subgoals in a learning hierarchy may be useful procedures in initial analyses of some instructional domains, these procedures do not provide a learning goal structure adequate for the design of learning progress maps or the identification of reference tasks within them.

2.4. The Case-Bereiter procedure for developmental instructional design has been used successfully and could with some modification serve as a framework for the identification and design of individual reference tasks for intermediate goals. It requires research to understand the performance strategies used by experts, and also by successful

and unsuccessful students, on each task.

2.5. The Campione-Brown procedure for the definition of transfer distances and tutoring hint hierarchies also should be useful in goal and reference task design. It also requires cognitive analysis of intermediate goals.

2.6. Simulation modeling of important instructional tasks can serve to specify intermediate learning goals and to suggest instructional or coaching strategies for them. Simulation models of reference tasks would help identify the cognitive structures and processes that need to be diagnosed in performance on such tasks and could also suggest their redesign for this purpose. These models also would be useful in evaluating the instruction that intervenes between reference tasks.

2.7. Two kinds of simulation models exist. One models the knowledge and skill required in goal performance. The other also models the learning process by which that knowledge and skill is acquired. The latter involves additional strong assumptions which may not be tenable in many instructional domains. Reference tasks should be conditioned on the first kind of model.

2.8. There are also computer-based tasks, often called "microworlds", which simulate aspects of the natural world whether or not they include simulations of aspects of human performance. These tasks may address important instructional goals, including some of the deeper or higher goals that are difficult to explicate. As such, they may be useful as reference tasks even though some redesign or augmentation is necessary for this purpose. Further, contrasts among existing microworld, intelligent tutoring, and other systems may offer radically alternative viewpoints for instructional goal and reference task definition. In work on electricity, for example, such a contrast might be that between the

approaches of Lesgold (1988), White and Frederiksen (1986), and Härtel (1987).

2.9. The mapping of tasks and intermediate goals into a curriculum space initially involves rough judgments of the difficulty or complexity of each and of proximity relations between them. The research already available on some tasks will be useful. But virtually no prior research has examined multiple alternative tasks of this sort, much less orderly arrays of such tasks. This then is a primary focus for SIIA research.

3. Designing reference tasks and probe tasks.

3.1. With the collection of existing tasks and models in hand as candidates and a provisional topographic mapping of these candidates onto a course curriculum, research would then focus on each task and the relations between them.

3.2. An early consideration is the appropriate grain size for the learning progress map, and this depends on the educational purposes the assessment is intended to serve. Different levels of educational policy and program evaluation demand different assessment models. Here we assume that the map is to cover a full course and be useful for classroom teachers in planning daily and weekly instruction. Thus, all reference tasks should have a grain size roughly like that of DAS and LAS, but probably cannot descend to the level of grain typically seen in CAS; some particular targets of difficulty in the domain, however, might require and justify this level. Instructional and assessment procedures such as those used in current DAS and LAS work may in some instances be computerizable as reference tasks. Candidate reference tasks coming from other research will also come in different grains. Redesign of some of these tasks to reach different grain sizes will often be required.

3.3. Since the psychological distances between some tasks may appear to be great, additional tasks will need to be chosen or created to fill gaps. These "probe" tasks

can be used to explore the ground between reference tasks and may become the basis for additional reference tasks. They may also suggest questions for the design of instruction to intervene between reference tasks. Existing achievement test items, used in constructed response format, may serve as probe tasks. Teacher and textbook questions and exercises may also serve this purpose.

3.4. Analysis of each reference task then proceeds by determining what constitutes and signifies different levels of performance on the task and the kinds of faulty performances it produces. Comparisons between advanced and beginning student performance, experimental manipulations of task characteristics, attempts at simulation, and correlations with other reference tasks and with other ability and achievement tests, provide empirical evidence regarding the construct interpretation of the target task.

3.5. A catalog of indicators in task performance is built up for each reference task. It is concerned particularly with symptoms of incomplete or faulty performance. Further cognitive analysis, perhaps using evidence from student interviews and think-aloud protocols, seeks to identify the sources of each type of error and the importance of each in undermining performance. The aim is to define interpretive constructs for different patterns of incomplete and faulty performance, and to design the reference tasks so as to distinguish the different interpretations clearly.

3.6. Reference tasks should be designed to provoke likely errors where they are expected. The catalogue of errors and types of faults built up in SIIA development should include indicators of weaknesses and inflexibilities as well as outright errors, and should allow for the observation of idiosyncratic "sloppy" errors where they occur. Errors of both commission and omission are a concern. Developing indicators of faulty structuring and its

remediation is probably the highest priority for research and diagnosis.

3.7. Learning progress involves the unlearning or restructuring of prior beliefs and habits as well as the learning of new ones. Reference tasks should be geared to detect this kind of transition when it occurs. But reference tasks should also allow for a diversity of alternative strategies and beliefs where these are constructive for learning progress.

3.8. Different fault interpretations should be associated with different next steps for instruction, so part of the interpretation-validation problem is to identify means by which each type of fault can be corrected with instruction.

3.9. Confidence bands need to be designed around alternative next step decisions, and a "harm" criterion should be applied in this design. Where the consequences to the student of incorrect interpretations or next step decisions are psychologically adverse or difficult to reverse, strong confidence is required.

3.10. Research needs to address the role of teachers in the interpretation of reference task performance and in the planning of intervening instruction. Reference tasks need to be made teacher-friendly. These issues should be addressed early rather than late in the design process

3.11. Guidelines should be developed to enable teachers to build their own probe tassk and auxilliary reference tasks to be added to the SIIA design. We can imagine the need for adaptation of existing reference tasks or the production of new curriculum modules that are locally relevant for teachers. Guidelines for such development or adaptation of reference tasks should be designed with sufficients flexibility so that teachers are able to use those guidelines to produce new curriculum modules and tasks that are locally appropriate.

3.12. Teachers should be directly involved in the development and use of SIIA

interpretation also because prior research suggests that teacher observation, of LAS sessions for example, enhances teacher expectations about learning progress for lower ability students.

4. Building diagnostic interpretations.

4.1. Individual reference tasks and coordinated series of such tasks should support interpretations of positive learning progress, as well as detect errors, faults, or other weaknesses that impede such progress at any particular point. Research on diagnosis thus proceeds on two levels: it connects next instructional steps with particular impediments to progress, but it also builds up an account of the psychology of learning progress across instructional steps. The fullest account will address both cognitive and conative aspects of learning and both general and domain specialized aspects of learning.

4.2. A diagnosis is a classification decision at both levels. At the level of next step decisions, diagnosis would seem to require validation by some variant of ATI methodology in each instance. The research needs to show that learners in different states of performance on a task are better served by the next instructional assignment they are given than they would be by some specified other assignment. But validation of a sequence of decisions considered as a whole may suffice in the early stages of SIIA development. The evaluation would need to show that this decision sequence results in expected learner progress whereas another does not.

4.3. Reference tasks ought to be geared to exhibit as well as promote the structures of deep understanding, the skills of thinking and reasoning and the efficient learning strategies in a domain, but also the motivational and self-regulatory functions that support and sustain learning progress. Each reference task design is a complex

measurement development and validation project.

4.4. Reference tasks ought also to exhibit the phase processes of learning and development. Accretion, structuring, and tuning, or knowledge acquisition, compilation, and automatization, or model progressions, or strategy incrementations, or various other types of conceptual shifts, may provide phase distinctions that task designs can address. They may also address certain phases of commitment, effort investment, intention protection, and action control. The designs should generate evidence useful in demonstrating and verifying such phase distinctions if they occur.

4.5. A crucial issue is transfer. Reference and probe tasks should incorporate transfer continua designed as in LAS to gauge the degree to which learning is welded to the tasks in which it occurred.

4.6. Three sources of diagnostic information should be distinguished. Readiness diagnosis involves information concerning the completeness of learning at whatever stage or phase a particular reference task represents; it signals advance to one of several next tasks. Remedial diagnosis involves information that signals specific faults in task performance and concerns what available remedial instructional steps are needed. Resistance diagnosis involves information suggesting general aptitudinal weaknesses that need attention in instructional design. Readiness diagnosis thus builds mainly on indicants of learning progress in the domain, remedial diagnosis on indicants of faulty learning in the domain, and resistance diagnosis on indicants of general or prior weaknesses from outside the domain. But the three are best used together to form interpretations.

4.7. We think that SIIA research should side-step the psychometric problems of change measurement at this juncture, as well as the definitional problems of "dynamic"

assessment. Individual reference tasks do not need to incorporate gain models. Learning progress across tasks can be described without explicit gain measures because qualitative as well as quantitative shifts in performance within and between tasks signify progress, and implicitly gain, without the need for explicit change computations. And learning is by definition "dynamic" regardless of how it is represented in score indices.

4.8. Learning progress will need to be defined differently in each domain. Each domain will have its interpretive constructs, based on its goal lattice structure and network of reference tasks. As a start, work in each domain might attempt to describe what kinds of performances constitute beginner, advanced beginner, competence, and advance competence levels. The research on various kinds of novice-expert comparisons may be helpful here, if it is not used to impose restrictions on the development of domain-specialized progress interpretations. In most domains, the progression from novice to expert is multivariate and probably nonlinear.

4.9. Diagnosis of conative aspects of learning progress will need to rely on questionnaire, interview, and informal observation in the early stages. But these aspects should not be ignored. It may be possible to design some reference tasks to provide performance-based conative indicators of progress (see Winne, in press; Zimmerman, 1990).

5. Evaluating learning progress systems.

5.1. Individual and class aggregate profiles can be obtained from learning progress maps. They serve to indicate further the targets of difficulty for individuals and groups to both teachers and designers, and they can thus guide further revision.

5.2. Formative evaluation, and adaptation to local conditions, is the sine qua

non of SIIA research and development. We do not imagine SIIA as being "designed" and then "implemented", as those terms are typically used. Rather, it is our hope but pehaps an unrealistic expectation that each such system will evolve in its time, place, and domain, as a function of continuous monitoring and tinkering, even though each may start from the rough common scheme described in this report (see Cronbach, 1982a, 1982b, 1986 for a justification of this view). Transferable principles may emerge from one site to be applied in another. But that will be an empirical question to be evaluated anew in each site, as well as each domain.

5.3. It follows that SIIA design should contain provisions for continuous monitoring and evaluation of its own functioning in each usage. It is unrealistic to expect that formal evaluation of a SIIA design can be conducted locally on a regular basis. But it may be possible to arrange regular, relatively automatic data entry for periodic review by a project team.

5.4. Field evaluations of SIIA should be planned to include measures of an assortment of relevant aptitude and achievement constructs. Both cognitive and conative constructs should be considered. In addition to aptitude measures reflecting prior knowledge and beliefs in the target domain, most evaluation studies will need to include measures of relevant tool skills, such as reading and numeric ability, and of fluid reasoning, achievement motivation, and anxiety. In addition to achievement measures reflecting depth of understanding and efficiency of skill, most studies will need to include measures of flexibility of knowledge use and transfer, capability for independent, self regulated learning, motivation for further learning, and self-concept as a learner. Aptitude constructs found to be important in continuing evaluation studies become candidates for inclusion in the

SIIA as part of resistance diagnosis.

6. Implementation.

6.1 Attention needs to be given to the critical role of the teacher and how SIIA will be implemented in actual classroom settings. The role of the teacher is likely to be altered greatly by the introduction and use of SIIA. Therefore, it is critical to involve teachers in the earliest stages of project development, as systems are conceptualized, developed, and implemented. Moreover, it is necessary to monitor and note the changing role of the teacher, throughout the implementation process (Mandinach & Cline, in press). Such observations can then feed into subsequent training and implementation activities.

6.2 Training needs to be provided to all participating teachers. The nature of the training should focus on the use and integration of the technology into existing curricula, the theoretical and practical constraints surrounding implementation, and use of the SIIA. Such training requires more than a focus on doing something different in the classroom. Rather, training here refers to working with and helping teachers to adapt and integrate the new perspectives into their instructional routines.

6.3 SIIA often require hardware and software that are unavailable in most schools. The developers must give prior thought to the appropriate provisions of equipment that will lead to a fair test of the potentials. Furthermore, the developers must acknowledge that in order for SIIA to have widespread impact, there is a need for use of hardware and software that are available and practical for implementation in classroom settings. Large in expensive computer systems will make SIIA unapproachable, impractical and unusable for most schools.

6.4 Just as the role of the teacher is likely to change as a result of SIIA, so too

will the nature of the classroom. Attention must be devoted to issues surrounding such evolving processes.

6.5 Accountability needs to be considered when implementing SIIA. The introduction of SIIA will likely cause a sequence of changes in expectations for learning outcomes and achievement, The changes are likely to be in opposition to be intransigent and long standing pressure for highly visible data on achievement and accountability. Working toward fundamental changes in the accountability system will be a lengthy process that must be acknowledged and dealt with as SIIA are introduced and implemented.

A Challenge

Integrated instructional and assessment systems arising from this work ought to be practically useful for instructional adaptation and for communication among students, teachers, parents, and others. But they also ought to be vehicles for research on the psychology of learning progress in school subject matters. Building domain topographies for instructional assessment of the sort envisioned in this report is a step toward theories of understanding for the domains addressed. Such topographies can bridge across many individual investigators by including key tasks from different lines of individual research. They can provide a network for the validation of particular interpretive constructs and measures. Most importantly, they represent a rich and thick description of what constitutes learning progress toward advanced competence in a domain. As research continues on each such system, this description should become more complete and more credible, even as it may also become more specialized for particular schools or student populations. In short, integrating assessment and instruction in this way appears to us to be the best route to theories of instruction in school domains.

References

Ackerman, P. (1989). Individual differences and skill acquisition. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), Learning and individual differences (pp. 165-217). New York: W. H. Freeman.

American Association for the Advancement of Science. (1989). Project 2061: Science for all Americans. Washington, DC: American Association for the Advancement of Science.

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

Anderson, J. R. (1988). The expert module. In M. C. Polson & J. J. Richardson (Eds.), Foundations of intelligent tutoring systems (pp. 21-53). Hillsdale, NJ: Erlbaum.

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. Science, 228, 456-462.

Anderson, J. R., Boyle, C. F., & Yost, G. (1985). The geometry tutor. In A. Joshi (Ed.), Proceedings of the Ninth International Joint Conference in Artificial Intelligence. Los Altos, CA: Morgan Kaufmann.

Anderson, J. R., Conrad, F. G., & Corbett, A. T. (1989). Skill acquisition and the LISP tutor. Cognitive Science, 13, 467-505.

Anderson, J. R., & Reiser B. (1984). The LISP tutor. Byte, 10, 159-175.

Atkinson, R. C. (1972). Ingredients for a theory of instruction. American Psychologist, 27, 921-923.

Baddeley, A. D. (1982). Your memory: A user's guide (1st American ed.). New York: Macmillan.

Baker, E. L., & O'Neil, H. F., Jr. (1987). Assessing instructional outcomes. In R. M. Gagné (Ed.), Instructional technology: Foundations (pp. 343-377). Hillsdale, NJ: Erlbaum.

Baron, J. B., & Sternberg, R. J. (Eds.). (1987). Teaching thinking skills: Theory and practice. New York: W. H. Freeman.

Barr, A., Beard, M., & Atkinson, R. C. (1976). The computer as a tutorial laboratory: The Stanford BIP project. International Journal of Man-Machine Studies, 8, 567-596.

Barr, A., & Feigenbaum, E. A. (Eds.). (1982). The handbook of artificial intelligence (Vol. 2). Los Altos, CA: William Kaufmann.

Bennett, R. E. (in press). Toward intelligent assessment: An integration of complex constructed response, artificial intelligence, and model-based measurement. In N. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum

Bennett, R. E., Gong, B., Kershaw, R. C., Rock, D. A., Soloway, E., & Macaladad, A. (1990). Assessment of an expert system's ability to grade and diagnose automatically students' constructed responses to computer science problems. In R. O. Freedle (Ed.), Artificial intelligence and the future of testing. Hillsdale, NJ: Erlbaum.

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

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (Eds.). (1956). Taxonomy of educational objectives. Handbook I: Cognitive domain. New York: David McKay.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). Handbook of formative and summative evaluation of student learning. New York: McGraw-Hill.

Bormuth, J. R. (1970). On the theory of achievement test items. Chicago, IL: University of Chicago Press.

Boyer, E. L. (1983). High school: A report on secondary education in America. New York: Harper & Row.

Boyle, C. F., & Anderson, J. R. (1984). Acquisition and automated instruction of geometry proof skills. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. Journal of Educational Measurement, 27, 93-108.

Braun, H. I., Bennett, R. E., Soloway, E., & Frye, D. (1989). Developing and evaluating a machine-scorable constrained constructed-response item. Princeton, NJ: Educational Testing Service.

Bricken, W. M. (1987). Analyzing errors in elementary mathematics. Unpublished doctoral dissertation, Stanford University, Stanford.

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: Erlbaum.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In P. H. Mussen (Ed.), Handbook of child psychology: Cognitive development (Vol. III, 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, J. S., Collins, A., & Duguid, P. (1989a). Debating the situation: A rejoinder to Palincsar and Wineberg. Educational Researcher, 18(4), 10-12, 62.

Brown, J. S., Collins, A., & Duguid, P. (1989b). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32-41.

Budoff, M. (1987a). Measures for assessing learning potential. In C. S. Lidz (Ed.), Dynamic assessment (pp. 173-195). New York: Guilford Press.

Budoff, M. (1987b). The validity of learning potential assessment. In C. S. Lidz (Ed.), Dynamic assessment (pp. 52-81). New York: Guilford Press.

Bunderson, C. V. (1969). Ability by treatment interactions in designing instruction for a hierarchial learning task. Paper presented to the American Educational Research Association.

Bunderson, C. V. (1973). The TICCIT project: Design strategy for educational innovation. In S. A. Harrison & L. M. Stolurow (Eds.), Productivity in higher education. Washington DC: National Institute of Education.

Bunderson, C. V., Inouye, D. K., & Olson, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 367-407). New York: Macmillan.

Burns, H. L., & Capps, C. G. (1988). Foundations of intelligent tutoring systems: An introduction. In M. C. Polson & J. J. Richardson (Eds.), Foundations of intelligent tutoring systems (pp. 1-19). Hillsdale, NJ: Erlbaum.

Burton, R. R. (1982). DEBUGGY: Diagnosis of errors in basic mathematical skills. In D. H. Sleeman & J. S. Brown (Eds.), Intelligent tutoring systems (pp. 157-183). New York: Academic Press.

Burton, R. R. (1988). The environment module of intelligent tutoring systems. In M. C. Polson & J. J. Richardson (Eds.), Foundations of intelligent tutoring systems (pp. 109-142). Hillsdale, NJ: Erlbaum.

Butterfield, E. C., Nielsen, D., Tangen, K. L., & Richardson, M. B. (1985). Theoretically based psychometric measures of inductive reasoning. In S. E. Embretson (Ed.) Test design: Developments in psychology and psychometrics (pp. 77-147). Orlando, FL: Academic Press.

Calfee, R. C. (1977). Assessment of independent reading skills: Basic research and practical applications. In A. S. Reber & D. L. Scarborough (Eds.), Toward a psychology of reading. Hillsdale, NJ: Erlbaum.

Calfee, R. C. (1981). Cognitive psychology and educational practice. Review of Research in Education, 9, 3-72.

Calfee, R. C. (1982) Cognitive models of reading: Implications for assessment and treatment of reading disability. In R. N. Malatesha & P. G. Aaron (Eds.), Reading disorders: Varieties and treatments. New York: Academic Press.

Calfee, R. C., & Spector, J. E. (1981). Separable processes in reading. In F. J. Pirozzolo & M. C. Wittrock (Eds.), Neuropsychological and cognitive processes in reading. New York: Academic Press.

Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), Dynamic assessment (pp. 82-115). New York: Guilford Press.

Campione, J. C., & Brown, A. L. (1990). Guided learning and transfer: Implications for approaches to assessment. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 141-172). Hillsdale, NJ: Erlbaum.

Campione, J. C., Brown, A. L., & Ferrara, R. A. (1982). Mental retardation and intelligence. In R. J. Sternberg (Ed.), Handbook of human intelligence (pp. 392-490). Cambridge: Cambridge University Press.

Carey, S. (1988). Reorganization of knowledge in the course of acquisition. In S. Strauss (Ed.), Ontogeny, phylogeny, and historical development. Norwood, NJ: Ablex.

Carr, B. (1977). WUSOR II: A computer aided instruction program with student modeling capabilities (AI Memo 417). Cambridge, MA: Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Carr, T. H. (1981). Building theories of reading ability: On the relation between individual differences in cognitive skills and reading comprehension. Cognition, 9, 73-114.

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

Case, R. (1985). Intellectual development: Birth to adulthood. Orlando, FL: Academic Press.

Case, R., & Bereiter, C. (1984). From behaviorism to cognitive behaviorism to cognitive development: Steps in the evolution of instructional design. Instructional Science, 13, 141-158.

Case, R., & Sandieson, R. (1988). A developmental approach to the identification and teaching of central conceptual structures in middle school science and mathematics. In M. Behr & J. Hiebert (Eds.) Research agenda in mathematics education: Number concepts and operations in the middle grades. Hillsdale, NJ: Erlbaum.

Case, R., Sandieson, R., & Dennis, S. (1986). Two cognitive developmental approaches to the design of remedial instruction. Cognitive Development, 1, 293-333.

Chant, V. G., & Atkinson, R. C. (1973). Optional allocation of instructional effort to interrelated learning strands. Journal of Mathematical Psychology, 10, 1-25.

Chi, M. T. H. (1978). Knowledge structures and memory development. In R. S. Siegler (Ed.), Children's thinking: what develops? (pp. 73-96). Hillsdale, NJ: Erlbaum.

Chi, M. T. H., Glaser, R., & Farr, M. J. (Eds.). (1988). The nature of expertise. Hillsdale, NJ: Erlbaum.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 1, pp. 7-75). Hillsdale, NJ: Erlbaum.

Chipman, S. F., Segal, J. W., & Glaser, R. (Eds.). (1985). Thinking and learning skills (Vol. 2). Hillsdale, NJ: Erlbaum.

Claeys, W., DeBoeck, P., Van Den Bosch, W., Biesmans, R., & Bohrer A. (no date). A comparison of one free format and two fixed format self-report personality assessment methods. Unpublished manuscript, Department of Psychology, University of Leuven, Belgium.

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In M. C. Wittrock (Ed.) Handbook of research on teaching (3rd ed., pp. 255-296). New York: Macmillan.

College Board Entrance Examination (1983). Academic preparation for college. New York: College Entrance Examination Board.

College Board Entrance Examination (1988). The College Board technical manual for the Advanced Placement Program. New York: College Board Entrance Examination.

Collins, A., & Brown, J. S. (1988). The computer as a tool for learning through reflection. In H. Mandl & A. Lesgold (Eds.), Learning issues for intelligence tutoring systems (pp. 1-18). New York: Springer-Verlag.

Corno, L., & Snow, R. E. (1986). Adapting teaching to individual differences among learners. In M. C. Wittrock (Ed.), Handbook of research on teaching (3rd. ed., pp. 605-629). New York: Macmillan.

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

Cronbach, L. J. (1975a). Five decades of public controversy over mental testing. American Psychologist, 30, 1-13.

Cronbach, L. J. (1975b). Beyond the two disciplines of scientific psychology. American Psychologist, 30, 116-127.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? New directions for testing and measurement - Measuring achievement over a decade - Proceedings of the 1979 ETS Invitational Conference. San Francisco: Jossey-Bass.

Cronbach, L. J. (1982a). Designing evaluations and social programs. San Francisco: Jossey-Bass.

Cronbach, L. J. (1982b). Prudent aspirations for social inquiry. In. W. H. Kruskal (Eds.), The social sciences: their nature and uses. Chicago: University of Chicago Press.

Cronbach, L. J. (1984). Essentials of psychological testing (4th ed.). New York: Harper & Row.

Cronbach, L. J. (1986). Social inquiry by and for earthlings. In D. W. Fiske & R. A. Shweder (Eds.), Metatheory in social science: Pluralities and subjectivities. Chicago: University of Chicago Press.

Cronbach, L. J. (1988a). Construct validation after thirty years. In R. Linn (Ed.), Intelligence: Measurement, theory, and public policy. Urbana, IL: University of Illinois Press.

Cronbach, L. J. (1988b). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Essentials of psychological testing (5th ed.). New York: Harper & Row.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"- or should we? Psychological Bulletin, 70, 68-80.

Cronbach, L. J., & Gleser, G. G. (1965). Psychological tests and personnel decisions. Urbana, IL: University of Illinois Press.

Cronbach, L. J., & Snow, R. E. (1977). Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington.

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

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. Exceptional Children, 52, 219-232.

Detterman, D. K., & Sternberg, R. J. (Eds.). (1982). How and how much can intelligence be increased. Norwood, NJ: Ablex.

Dreyfus, H. L., & Dreyfus, S. E. (1986). Mind over machine. New York: Free Press.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. Psychological Review, 95, 256-273.

Educational Technology Center (1988). Making sense of the future: A position paper on the role of technology in science, mathematics, and computer education (Position Paper PP 88-0). Cambridge, MA: Harvard University Graduate School of Education, Educational Technology Center.

Embretson, S. E. (Ed.). (1985). Test design: Developments in psychology and psychometrics. Orlando, FL: Academic Press.

Embretson, S. E. (1987). Toward development of a psychometric approach. In C. S. Lidz (Ed.), Dynamic assessment (pp. 141-170). New York: Guilford Press.

Embretson, S. E. (1990). Diagnostic testing by measuring learning processes: Psychometric considerations for dynamic testing. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 407-432). Hillsdale, NJ: Erlbaum.

Emmerich, W. (1989). Appraising the cognitive features of subject tests. (Research Report RR 89-53). Princeton, NJ: Educational Testing Service.

Entwistle, N. (1987). Explaining individual differences in school learning. In E. DeCorte, H. Lodewijks, R. Parmentier, & P. Span (Eds.), Learning and instruction: European research in an international context (Vol. 1). Leuven, Belgium and Oxford, UK: Leuven University Press and Pergamon Press.

Feuerstein, R. (1979). The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques. Baltimore: University Park Press.

Feuerstein, R., Jensen, M., Hoffman, M. B., & Rand, Y. (1985). Instrumental enrichment: An intervention program for structural modifiability: Theory and practice. In J. W. Segal, S. F. Chipman, & R. Glaser (Eds.), Thinking and learning skills (Vol. 1, pp. 43-82). Hillsdale, NJ: Erlbaum.

Feuerstein, R., Rand, Y., Jensen, M. R., Kaniel, S., & Tzuriel, D. (1987). Prerequisites for assessment of learning potential: The LPAD model. In C. S. Lidz (Ed.), Dynamic assessment (pp. 35-81). New York: Guilford Press.

Forehand, G. A., & Bunderson, C. V. (1987a). Basic concepts of mastery assessment systems. Princeton, NJ: Educational Testing Service.

Forehand, G. A., & Bunderson, C. V. (1987b). Mastery assessment systems and educational objectives. Princeton, NJ: Educational Testing Service.

Forehand, G. A., & Rice, M. W. (1988). Guides to learning and instruction: Final report 1987-88 formative research project: Volume 1. Princeton, NJ: Educational Testing Service.

Frase, L. T., & Diel, M. (1986). UNIX Writer's Workbench: Software for streamline communication. T.H.E. Journal, 14(3), 74-78.

Frederiksen, J. R. (1982). A componential theory of reading skills and their interactions. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol.1, pp. 125-180). Hillsdale, NJ: Erlbaum.

Frederiksen, J., & White, B. (1990). Intelligent tutors as intelligent testers. In N. Frederiksen, R. Glaser, A. Lesgold, &. M. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 1-25). Hillsdale, NJ: Erlbaum.

Frederiksen, J. R., White, B. Y., Collins, A., & Eggan, G. (1988). Intelligent tutoring systems for electronic troubleshooting. In J. Psotka, L. D. Massey, & S. A. Mutter (Eds.), Intelligent tutoring systems: Lessons learned (pp. 351-368). Hillsdale, NJ: Erlbaum.

Frederiksen, N. (1984). The real test bias: Influences on teaching and learning. American Psychologist, 39, 193-202.

Frederiksen, N., Glaser, R., Lesgold, A., & Shafto, M. (Eds.). (1990). Diagnostic monitoring of skill and knowledge acquisition. Hillsdale, NJ: Erlbaum.

Frederiksen N., Mislevy, R., & Bejar, I. (in press). Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum.

Fuchs, L. S., & Fuchs, D. (1990). Curriculum-based assessment. In C. Reynolds & R. R. Kamphaus (Eds.), Handbook of psychological and educational assessment of children (Vol. 1): Intelligence and achievement. New York: Guilford Press.

Gagné, R. M. (1965). The conditions of learning (1st ed.). New York: Holt, Rinehart & Winston.

Gagné, R. M. (1968). Learning hierarchies. Educational Psychologist, 6, 1-9.

Gagné, R. M. (Ed.). (1987). Instructional technology: Foundations. Hillsdale, NJ: Erlbaum.

Gagné, R. M., Briggs, L. J., & Wager, W. W. (1988). Principles on instructional design (3rd ed.). New York: Holt, Rinehart and Winston.

Gardner, H. (1983). Frames of mind. New York: Basic Books

Gardner, H., & Hatch, T. (1989). Multiple intelligences go to school. Educational Researcher, 18(8), 4-10.

Gentner, D., & Stevens, A. L. (Eds.). (1983). Mental models. Hillsdale, NJ: Erlbaum.

Gitomer, D. H. (1988). Individual differences in technical troubleshooting. Human Performance, 1, 111-131.

Gitomer, D. H. (1989, November) Developing a portfolio culture that enables learners. Paper presented at the 1989 National Summit Conference on the Arts and Education, Washington, DC.

Gitomer, D. H., & Van Slyke, D. A. (1987/88). Error analysis and tutor design. Machine-Mediated Learning, 2, 333-350.

Glaser, R. (1963). Instructional technology and the measurement of learning systems. American Psychologist, 18, 510-522.

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

Glaser, R. (1985). Thoughts on expertise (Tech. Rep. No. 8). Pittsburgh: University of Pittsburgh, Learning Research and Development Center.

Glaser, R., & Bassok, M. (1989). Learning theory and the study of instruction. Annual Review of Psychology, 40, 631-666.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. R. Ronning, J. G. Glover, J. C. Conoley, & J. C. Witt (Eds.), The influence of cognitive psychology in testing (pp. 41-85). Hillsdale, NJ: Erlbaum.

Goldstein, I. (1979). The genetic graph: A representation for the evolution of procedural knowledge. International Journal of Man-Machine Studies, 11, 51-77.

Goldstein, I., & Carr, B. (1977). The computer as coach.: An athletic paradigm for intellectual education. Proceeding of 1977 Annual Conference of the Association for Computing Machinery, Seattle, 227-233.

Gong, B. (1988). "Mastery maps" and mastery assessment: An initial description and examples. Unpublished manuscript, Princeton, NJ, Educational Testing Service.

Gray, L. E. (1982). Aptitude constructs, learning processes, and achievement. Unpublished report, Stanford University.

Greeno, J. G. (1976). Cognitive objectives of instruction: Theory of knowledge for solving problems and answering questions. In D. Klahr (Ed.), Cognition and instruction (pp. 123-159). Hillsdale, NJ: Erlbaum.

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

Greeno, J. G. (1980). Some examples of cognitive task analysis with instructional implications. In R. E. Snow, P-A. Federico, & W. E. Montague (Eds.), Aptitude, learning, and instruction: Vol. 2. Cognitive process analyses of learning and problem solving (pp. 1-21). Hillsdale, NJ: Erlbaum.

Greeno, J. G. (1986a). Advancing cognitive science through development of advanced instructional systems. Machine-Mediated Learning, 1, 327-343.

Greeno, J. G. (1986b, April). Mathematical cognition: Accomplishments and challenges in research. Invited address to the American Educational Research Association, San Francisco.

Greeno, J. G. (1988). Situations, mental models, and generative knowledge. In D. Klahr & K. Kotovsky (Eds.). Complex information processing: The Impact of Herbert A. Simon. Hillsdale, NJ: Erlbaum.

Greeno, J. G., & Simon, H. A. (1988). Problem solving and reasoning. In R. C. Atkinson, R. Hernstein, G. Lindzey, & R. D. Luce (Eds.), Steven's handbook of experimental psychology (rev. ed.). New York: Wiley.

Guilford, J. P. (1967). The nature of human intelligence. New York: McGraw-Hill.

Haertel, E., & Calfee, R. C. (1983). School achievement: Thinking about what to test. Journal of Educational Measurement, 20, 119-132.

Härtel, H. (1987). A qualitative approach to electricity (IRL Report No. IRL87-0001). Palo Alto, CA: Institute for Research on Learning.

Halff, H. M. (1988). Curriculum and instruction in automated tutors. In M. C. Polson & J. J. Richardson (Eds.), Foundations of intelligent tutoring systems (pp. 79-108). Hillsdale, NJ: Erlbaum.

Heller, J. I., & Greeno, J. G. (1979). Information processing analyses of mathematical problem solving. In R. W. Tyler & S. H. White (Eds.), Testing, teaching and learning. Washington, DC: National Institute of Education.

Howard, K. (1990, Spring) Making the writing portfolios real. The Quarterly of the National Writing Project and the Center for the Study of Writing, 12(2), 4-7, 27.

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

Johnson, L., & Soloway, E. (1984). PROUST: Knowledge-based program debugging. Proceedings of the Seventh International Software Engineering Conference, 369-380.

Johnson, P. H. (1983). Reading comprehension assessment: A cognitive basis. Newark, DE: International Reading Association.

Johnson, W. B. (1988). Pragmatic instructions in research, development, and implementation of intelligent tutoring systems. In M. C. Polson & J. J. Richardson (Eds.), Foundations of intelligent tutoring systems (pp. 191-207). Hillsdale, NJ: Erlbaum.

Johnson-Laird, P. N. (1983). Mental models. Cambridge: Cambridge University Press.

Judd, C. H. (1915). Psychology of high school subjects. Boston: Ginn.

Judd, C. H. (1936). Education as cultivation of the higher mental processes. New York: Macmillan.

Kanfer, R., & Ackerman, P. L. (in press). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. Journal of Applied Psychology.

Kanfer, R., Ackerman, P. L., & Cudeck, R. (Eds.). (1989). Abilities, motivation, and methodology: The Minnesota Symposium on Learning and Individual Differences. Hillsdale, NJ: Erlbaum.

Kirsch, I. S. (1987, September). Measuring adult literacy. In Towards defining literacy. Symposium conducted at the National Advisory Council in Adult Education, Literacy Research Center, University of Pennsylvania, Philadelphia.

Kirsch, I. S., & Jungeblut, A. (1986). Literacy: Profiles of America's young adults--final report (Report No. 16-PL-02). Princeton, NJ: National Assessment of Educational Progress.

Kirsch, I. S., & Mosenthal, P. B. (1988). Understanding document literacy: Variables underlying the performance of young adults (Research Report RR-88-62). Princeton, NJ: Educational Testing Service.

Kirsch, I. S., Mosenthal, P. B., & Rock, D. A. (1988). The influence of reading patterns on the proficiency of young adults (Research Report RR-88-61). Princeton, NJ: Educational Testing Service.

Kitcher, P. (1984). The nature of mathematical knowledge. New York: Oxford University Press.

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

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). Taxonomy of educational objectives. Handbook II: Affective domain. New York: McKay.

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

Kuhn, T. S. (1970). The structure of scientific revolutions (2nd ed.). Chicago: University of Chicago Press.

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

Kyllonen, P. C., & Shute, V. J. (1989). A taxonomy of learning skills. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), Learning and individual differences (pp. 117-163). New York: W. H. Freeman.

Lajoie, S. P., & Lesgold, A. (1989). Apprenticeship training in the workplace: Computer coached practice environment as a new form of apprenticeship. Machine-Mediated Learning, 3, 7-28.

Lampert, M. (1986). Knowing, doing, and teaching multiplication. Cognition and Instruction, 3, 305-342.

Lampert M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. American Educational Research Journal, 27, 29-63.

Larkin, J., Reif, F., Carbonell, J., & Gugliotte, A. (1985). FERMI: A flexible expert reasoner with multi-domain inferencing. Cognitive Science, 12, 101-138.

Lepper, M. R., & Chabay, R. W. (1985). Intrinsic motivation and instruction: Conflicting views on the role of motivational processes in computer-based education. Educational Psychologist, 20, 217-230.

Lepper, M. R., & Malone, T. W. (1987). Intrinsic motivation and instructional effectiveness in computer-based education. In R. E. Snow & M. C. Farr (Eds.), Aptitude, learning and instruction: Vol. 3. Conative and affective process analysis (pp. 255-286). Hillsdale, NJ: Erlbaum.

Lesgold, A. (1988). Toward a theory of curriculum for use in designing intelligent instructional systems. In H. Mandl & A. Lesgold (Eds.), Learning issues for intelligent tutoring systems (pp. 114-137). New York: Springer-Verlag.

Lesgold, A., Bonar, J., & Ivill, J. (1987). Toward intelligent systems for testing (Tech. Rep. No. LSP-1). Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Lesgold, A., Lajoie, S., Bunzo, M., & Eggan, G. (in press). SHERLOCK: A coached practice environment for an electronics troubleshooting job. In J. Larkin, R. Chabay, & C. Scheftic (Eds.), Computer assisted instruction and intelligent tutoring systems: Establishing communication and collaboration. Hillsdale, NJ: Erlbaum.

Lewis, M. (1989, March). Developing and evaluating the CMU algebra tutor: Tension between theoretically and pragmatically driven design. Paper presented at American Educational Research Association, San Francisco.

Lidz, C. S. (Ed.). (1987). Dynamic assessment. New York: Guilford Press.

Littman, D., & Soloway, E. (1988). Evaluating ITSs: The cognitive science perspective. In M. C. Polson & J. J. Richardson (Eds.), Foundations of intelligent tutoring systems (pp. 209-242). Hillsdale, NJ: Erlbaum.

Lord, F. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Macdonald, N. H., Frase, L. T., Gingrich, P. S., & Keenan, S. A. (1982). The Writer's Workbench: Computer aids for text analysis. IEEE Transactions on Communications, 30(1), 105-110.

Malone, T. W., & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. In R. E. Snow & M. C. Farr (Eds.), Aptitude, learning, and instruction: Vol. 3. Conative and affective process analysis (pp. 223-253). Hillsdale, NJ: Erlbaum.

Mandinach, E. B. (1984). The role of strategic planning and self-regulation in learning an intellectual computer game. Unpublished doctoral dissertation, Stanford University, Stanford.

Mandinach, E. B. (1987). Clarifying the "A" in CAI for learners of different abilities. Journal of Educational Computing Research, 3, 113-128.

Mandinach, E. B., & Cline, H. F. (in press). Implementing technology-based learning environments: Systems thinking and curriculum innovation. Hillsdale, NJ: Erlbaum.

Mandl, H., & Lesgold, A. (Eds.). (1988). Learning issues for intelligent tutoring systems. New York: Springer-Verlag.

Mandler, J. M. (1984). Stories, scripts, and scenes: Aspects of schematic theory. Hillsdale, NJ: Erlbaum.

Marshall, S. P. (1988). Assessing schema knowledge (Technical Report). San Diego, CA: San Diego State University, Center for Research in Mathematics and Science Education.

Marshall, S. P. (1990). Generating good items for diagnostic tests. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.). Diagnostic monitoring of skill and knowledge acquisition (pp. 433-452). Hillsdale, NJ: Erlbaum.

Martinez, M. E., & Lahart, C. M. (1990). Profile: Student characteristics from the 1986 and 1988 NAEP Assessments (RR-90-20). Princeton, NJ, Educational Testing Service.

Marton, F. (1981). Phenomenography: Describing conceptions of the world around us. Instructional Science, 10, 177-200.

Marton, F. (1983). Beyond individual differences. Educational Psychology, 3, 291-305.

Marton, F., Hounsell, D. S., & Entwistle, N. J. (1984). The experience of learning. Edinburgh: Scottish Academic Press.

McArthur, D., & Stasz, C. (1989). An intelligent tutor for basic algebra (WD-2781-NSF). Santa Monica, CA: The Rand Corporation.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. Stevens (Eds.). Mental models (pp. 299-324). Hillsdale, NJ: Erlbaum.

McCloskey, M., & Kargon, R. (1988). The meaning and use of historical models in the study of intuitive physics. In S. Strauss (Ed.), Ontogeny, phylogeny, and historical development. Norwood, NJ: Ablex.

McK. .chie, W. J., Pintrich, P. R., Lin, Y-G. (1985). Teaching learning strategies. Educational Psychologist, 20, 153-160.

Melton, A. W. (Ed.). (1964). Categories of human learning. New York: Academic Press.

Merrill, M. D. (1983). Component display theory. In C. M. Reigeluth (Ed.), Instructional design theories and models (pp. 279-333). Hillsdale, NJ: Erlbaum.

Merrill, M. D. (1987). A lesson based on component display theory. In C. M. Reigeluth (Ed.), Instructional theories in action (pp. 201-244). Hillsdale, NJ: Erlbaum.

Merrill, M. D., & Boutwell, R. C. (1973). Instructional development: Methodology and research. In F. N. Kerlinger (Ed.), Review of research in education (Vol. 1). Itasca, IL: Peacock.

Messick, S. J. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-103). New York: Macmillan.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 335-366). New York: Macmillan.

Miller, J. R. (1988). The role of human-computer interaction in intelligent tutoring systems. In M. C. Polson & J. J. Richardson (Eds.), Foundations of intelligent tutoring systems (pp. 143-189). Hillsdale, NJ: Erlbaum.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects follow different solution strategies. Psychometrika. 55(2), 195-215.

Murname, R. J., & Raizen, S. A. (1988). Improving indicators of the quality of science and mathematics education in grades K-12. Washington, DC: National Academy Press.

Murphy, J., & Bunderson, C. V. (1988). An update on the concepts and status of mastery assessment systems. Princeton, NJ: Educational Testing Service.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington, DC: U. S. Department of Education.

National Council of Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: National Council of Teachers of Mathematics.

Naveh-Benjamin, M., McKeachie, W. J., Lin, Y-G., & Tucker, D. G. (1986). Inferring students' cognitive structures and their development using the "Ordered Tree Technique." Journal of Educational Psychology, 78, 130-140.

Neches, R. C. (1982). Simulation systems for cognitive psychology. Behavior Research Methods & Instrumentation, 14, 77-91.

Nickerson, R. S., Perkins, D. N., & Smith, E. E. (1985). The teaching of thinking. Hillsdale, NJ: Erlbaum.

Nitko, A. J., (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 447-474). New York: Macmillan.

Norman, D. (1982). Learning and memory. San Francisco: W. H. Freeman.

Norman, D. (1982). Some observations on mental models. In D. Gentner & A. Stevens (Eds.), Mental models (pp. 7-14). Hillsdale, NJ: Erlbaum.

Ohlsson, S. (1984a, June). Attentional heuristics in human thinking. Proceedings of the Sixth Conference of the Cognitive Science Society, Boulder, CO.

Ohlsson, S. (1984b). Induced strategy shifts in spatial reasoning. Acta Psychologica, 57, 47-67.

O'Neil, H. F. Jr., (Ed.). (1978). Learning strategies. New York: Academic Press.

O'Neil, H. F. Jr., & Spielberger, C. D. (Eds.). (1979). Cognitive and affective learning strategies. New York: Academic Press.

Palincsar, A. S. (1989). Less charted waters. Educational Researcher, 18(4), 5-7.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and monitoring activities. Cognition and Instruction, 1, 117-175.

Pask, G. (1976). Styles and strategies of learning. British Journal of Educational Psychology, 46, 128-148.

Perkins, D. N., & Martin, F. (1985). Fragile knowledge and neglected strategies in novice programmers (TR85-22). Cambridge: Harvard University, Educational Technology Center.

Perkins, D. N., Martin, F., & Farady, M. (1986). Loci of difficulty in learning to program (TR86-6). Cambridge: Harvard University, Educational Technology Center.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? Educational Researcher, 18(1), 16-25.

Pines, A. L., Novak, J. D., Posner, G. J., & Van Kirk, J. (1978). The clinical interview: A method for evaluating cognitive structure (Research Report No. 6). Ithaca, NY: Cornell University Press.

Polson, P. G., & Richardson, J. J. (Eds.). (1988). Foundations of intelligent tutoring systems. Hillsdale, NJ: Erlbaum.

Popham, W. J. (1975). Educational evaluation. Englewood Cliffs, NJ: Prentice Hall.

Psotka, J., Massey, L. D., & Mutter, S. A. (Eds.). (1988). Intelligent tutoring systems: Lessons learned. Hillsdale, NJ: Erlbaum.

Reigeluth, C. M. (Ed.). (1983). Instructional design theories and models: An overview of their current status. Hillsdale, NJ: Erlbaum.

Reigeluth, C. M. (Ed.). (1987a). Instructional theories in action: Lessons illustrating selected theories and models. Hillsdale, NJ: Erlbaum.

Reigeluth, C. M. (Ed.). (1987b). Lesson blueprints based on the elaboration theory of instruction. In C. M. Reigeluth (Ed.), Instructional design theories and models (pp. 245-288). Hillsdale, NJ: Erlbaum.

Reigeluth, C. M., & Stein, F. S. (1983). The elaboration theory of instruction. In C. M. Reigeluth (Ed.), Instructional design theories and models. Hillsdale, NJ: Erlbaum.

Reiser, B., Anderson, J., & Farrell, R. (1985). Dynamic student modeling in an intelligent tutor for LISP programming. Proceedings of the Ninth International Joint Conference in Artificial Intelligence, 8-14.

Resnick, L. B. (1976). Task-analysis in instructional design: Some cases from mathematics. In D. Klahr (Ed.), Cognition and instruction (pp. 51-80). Hillsdale, NJ: Erlbaum.

Resnick, L. B. (1987). Education and learning to think. Washington: National Academy Press.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in mathematics. In H. P. Ginsberg (Ed.), Development of mathematical thinking (pp. 153-196). New York: Academic Press.

Rock, D. A., & Pollack-Ohls, J. (1987). Measuring gains--A new look at an old problem. Princeton, NJ: Educational Testing Service.

Rogoff, B., & Lave, J. (Eds.). (1984). Everyday cognition: Its development in social context. Cambridge, MA: Harvard University Press.

Rogosa, D. R., Brandt, D., & Zimkowski, M. (1982). A growth curve approach to the measurement of change. Psychological Bulletin, 90, 726-748.

Rummelhart, D. E., & Norman, D. A. (1978). Accretion, tuning, and restructuring: Three modes of learning. In J. W. Cotton & R. Klatzky (Eds.), Semantic factors in cognition. Hillsdale, NJ: Erlbaum.

Ryans, D. G. (1963). An information systems approach to education. Santa Monica, CA: Systems Development Corp.

Scandura, J. M. (1983). Instructional strategies based on the structural learning theory. In C. M. Reigeluth (Ed.), Instructional design theories and models. Hillsdale, NJ: Erlbaum.

Schoenfeld, A. H. (1985). Mathematical problem solving. Orlando: Academic Press.

Schwab, J. J. (1962). The concept of the structure of the discipline. Educational Record, 43, 197-205.

Schwab, J. J. (1978). Science, curriculum and liberal education: Selected essays. Chicago: University of Chicago Press.

Schwartz, S. (1984). Measuring reading competence: A theoretical-prescriptive approach. New York: Plenum Press.

Segal, J. W., Chipman, S. F., & Glaser, R. (Eds.). (1985). Thinking and learning skills (Vol. 1). Hillsdale, NJ: Erlbaum.

Shavelson, R. J., Pine, J., Goldman, S. R., Baxter, G. P., & Hine, M. S. (1989, June). New technologies for assessing science achievement. Paper presented at the American Psychological Society, Washington, DC.

Sheehan, K., & Mislevy, R. J. (1989). Integrating cognitive and psychometric models to measure document literacy (Research Report RR-89-51-ONR). Princeton, NJ: Educational Testing Service.

Shulman, L. S., & Ringstaff, C. (1989). Current research on the psychology of learning and tutoring. In A. Bork & H. Weinstock (Eds.). Designing computer-based learning material. New York: Springer-Verlag.

Shute, V. (1989, March). Individual differences in learning from an intelligent tutoring system. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Shute, V. (1990, April). A comparison of two computer-based learning enviroments. In R. Lehrer (Chair), Computers and cognitive tools. Symposium conducted at the meeting of the American Educational Research Association, Boston.

Shute, V. J., Glaser, R., & Rughaven, K. (1989). Inference and discovery in an exploratory laboratory. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), Learning and individual differences (pp. 279-326). New York: Freeman.

Siegler, R. S. (1976). Three aspects of cognitive development. Cognitive Psychology, 8, 481-520.

Siegler, R. S. (1978). The origins of scientific reasoning. In R. S. Siegler (Ed.), Children's thinking: What develops? Hillsdale, NJ: Erlbaum.

Siegler, R. S. (1986). Children's thinking. Englewood Cliffs, NJ: Prentice-Hall.

Siegler, R. S., & Campbell, J. (1989). Individual differences in children's strategy choices. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), Learning and individual differences (pp. 219-254). New York: W. H. Freeman.

Siegler, R. S., & Campbell, J. (1990). Diagnosing individual differences in strategy choice procedures. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 113-139). Hillsdale, NJ: Erlbaum.

Simon, H. A. (1976). Identifying basic abilities underlying intelligent performance of complex tasks. In L. B. Resnick (Ed.), The nature of human intelligence (pp. 65-98). Hillsdale, NJ: Erlbaum.

Simon, H. A. (1978). Information-processing theory of human problem solving. In W. K. Estes (Ed.), Handbook of learning and cognitive processes: (Vol. 5): Human information processing (pp. 271-295). Hillsdale, NJ: Erlbaum.

Sleeman, D. (1987). PIXIE A shell for developing intelligent tutoring systems. In R. W. Lawler & M. Yazdani (Eds.), Artificial intelligence and education (Vol. 1, pp. 239-265). Norwood, NJ: Ablex.

Snelbecker, G. E. (1985). Learning theory, instructional theory and psychoeducational design. Lanham, MD: University Press of America.

Snow, R. E. (1972). Individual differences in learning-related processes. Paper presented at American Educational Research Association, Chicago.

Snow, R. E. (1978). Eye fixation and strategy analysis of individual differences in cognitive aptitudes. In A. M. Lesgold, J. W. Pellegrino, S. D. Fokkema, & R. Glaser (Eds.), Cognitive psychology and instruction (pp. 299-308). New York: Plenum.

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

Snow, R. E. (1981). Toward a theory of aptitude learning: Fluid and crystallized abilities and their correlates. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), Intelligence and learning (pp. 345-362). New York: Plenum.

Snow, R. E. (1982). Training of intellectual aptitude: In D. K. Detterman & R. J. Sternberg (Eds.), How and how much can intelligence be increased (pp. 1-37). Norwood, NJ: Ablex.

Snow, R. E. (1987). Aptitude complexes in education. In R. E. Snow & M. C. Farr (Eds.), Aptitude, learning, and instruction: Vol. 3. Conative and affective process analysis (pp. 11-34). Hillsdale, NJ: Erlbaum.

Snow, R. E. (1988). Progress in measurement, cognitive science, and technology that can change the relation between instruction and assessment. In E. E. Freeman (Ed.) Assessment in the service of learning: Proceedings of the 1987 ETS Invitational Conference (pp. 9-25). Princeton, NJ: Educational Testing Service.

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

Snow, R. E. (1989b). Cognitive-conative aptitude interactions in learning. In R. Kanfer, P. L. Ackerman, & R. Cudeck, (Eds.), Abilities, motivation, and methodology (pp. 435-474). Hillsdale, NJ: Erlbaum.

Snow, R. E. (1989c). Toward assessment of cognitive and conative structures in learning. Educational Researcher, 18(9), 8-14.

Snow, R. E. (in press). The concept of aptitude. In R. E. Snow & D. F. Wiley, (Eds.), Improving inquiry in social science.

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

Snow, R. E., & Lohman, D. F. (1984). Toward a theory of cognitive aptitude for learning from instruction. Journal of Educational Psychology, 76, 347-376.

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

Snow, R. E., Wescourt, K., & Collins, J. (1979). Individual differences in aptitude and learning from interactive computer-based instruction (Tech. Rep. No. 10). Stanford, CA: Stanford University, Aptitude Research Project, School of Education.

Sternberg, R. J. (1982). Reasoning, problem solving, and intelligence. In R. J. Sternberg (Ed.), Handbook of human intelligence (pp. 225-307). Cambridge: Cambridge University Press.

Sternberg, R. J. (1985). Beyond IQ. New York: Cambridge University Press.

Sternberg, R. J. (1986). Understanding and increasing your intelligence. San Diego: Harcourt, Brace, Jovanovich.

Stevens, G. H., & Scandura, J. M. (1987). A lesson design based on instructional prescriptions from the structured learning theory. In C. M. Reigeluth (Ed.), Instructional theories in action (pp. 161-180). Hillsdale, NJ: Erlbaum.

Strauss, S. (Ed.). (1988). Ontogeny, phylogeny, and historical development. Norwood, NJ: Ablex.

Swanson, J. (1989). One-to-one tutoring: A experimental evaluation of effective tutoring strategies. Unpublished report, School of Education, Stanford University.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error analysis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 453-488). Hillsdale, NJ: Erlbaum.

Thurstone, L. L. (1947). Multiple factor analysis. Chicago: University of Chicago Press.

U. S. Department of Education, (1984). The nation responds: Recent efforts to improved education. Washington, DC: US Government Printing Office.

Vanderlinden, W. J. (1981). Using aptitude measurements for the optimal assignment of subjects to treatments with and without mastery scores. Psychometrika, 46, 257-274.

Van Lehn, K. C. (1988). Student modeling. In M. C. Polson & J. J. Richardson (Eds.), Foundations of intelligent tutoring systems (pp. 55-78). Hillsdale, NJ: Erlbaum.

Vernon, P. E., (1950). The structure of human abilities. New York: Wiley.

Vosniadou, S., & Brewer W. F. (1987). Theories of knowledge restructuring in development. Review of Educational Research, 57, 51-67.

Vye, N. J., Burns, S., Declos, V. R., & Bransford, J. D. (1987). A comprehensive approach to assessing intellectually handicapped children. In C. S. Lidz (Ed.), Dynamic assessment (pp. 327-359). New York: Guilford Press.

Vygotsky, L. S. (1978). Mind in society. Cambridge, MA: Harvard University Press.

Webb, N. M. (1982). Group composition, group interaction, and achievement in cooperative small groups. Journal of Educational Psychology, 74, 475-482.

Weinstein, C. E., Goetz, E. T., & Alexander, P. A. (Eds.). (1988). Learning and study strategies: Issues in assessment, instruction, and evaluation. New York: Academic Press.

Wenger, E. (1987). Artificial intelligence and tutoring systems. Los Altos, CA: Morgan Kaufmann.

White, B., & Frederiksen, J. (1986). Progressions of qualitative models as a foundation for intelligent learning environments (Rep. No. 6277). Cambridge, MA: BBN Laboratories.

White, B., & Frederiksen, J. (1987). Qualitative models and intelligent learning environments. In R. Lawler & M. Yazdani (Eds.), Artificial intelligence and education (Vol. 1, pp. 281-305). New Ycrk: Ablex.

White, B., & Horwitz, P. (1987). Thinker tools: Enabling children to understand physical laws (BBN Report No. 6470). Cambridge, MA: BBN Laboratories.

Wiley D. E., & Haertel, E. H. (in press). Models of skill patterns and sequences. In N. Frederiksen, R. Mislevy, & I. Bejar (Eds.), Test theory for a new generation on tests. Hillsdale, NJ: Erlbaum.

Wilson, S. M. (1988). Understanding historical understanding: Subject matter knowledge and the teaching of U. S. history. Unpublished doctoral dissertation, Stanford University, Stanford.

Wineberg, S. S. (1989). Remembrance of theories past. Educational Researcher, 18(4), 7-10.

Wineberg, S. S. (1990). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary evidence. Unpublished doctoral dissertation, Stanford University, Stanford.

Winne, P. H. (in press). State-of-the-art instructional computing systems that offend instruction and bootstrap research. In P. H. Winne & M. Jones (Eds.), Foundations and frontiers in instructional computing systems. New York: Springer-Verlag.

Winne, P. H., & Jones, M. (Eds.). (in press). Foundations and frontiers in instructional computing systems. New York: Springer-Verlag.

Wiser, M. (1988). The differentiation of heat and temperature: History of science and expert-novice shift. In S. Strauss (Ed.), Ontogeny, phylogeny, and historical development. Norwood, NJ: Ablex.

Zimmerman, B. J. (Ed.) (1990). Self-regulated learning and academic achievement. [Special issue]. Educational Psychologist, 25(1).

Table 1

Basic Issues for Research and Development

1. How are instructional goals, domains, and treatments defined?
   - Does the system assume hierarchically organized goals?
   - Does it accommodate multiple export viewpoints?
   - Is it limited to well-structured domains?
   - Are instructional treatments explicit?
   - At what grain-size does the system operate?

2. How are assessment tasks, tests, and other indicators of learning progress designed and used?
   - Do assessments arise from learning tasks?
   - Do assessments provide adaptive instructional decisions?
   - Are the natural products of learning used in assessment?

3. What roles are envisioned for teachers?
   - Does the system fit within conventional classroom procedures?
   - Is the design teacher-friendly?
   - Does the system use teacher observation and judgment as part of assessment?
   - Is teacher training provided for?
   - Are communication needs beyond the teacher-student relationship considered?

4. What theory of learning, development, and transfer is assumed?
   - How is learning progress characterized?
   - Is provision made for transfer assessment?
   - How are errors treated in instruction?
   - What theory of expertise is assumed?
   - What use is made of individual difference among students arising from outside the system?...inside the system?

5. How is diagnosis defined and treated?
   - What is diagnosed?
   - How is diagnosis achieved?
   - For what purposes is diagnosis used?
   - How is diagnosis validated?

6. How is the system evaluated?
   - Is evaluation built into the design?
   - Is the system self adaptive?
   - Are there provisions for evaluation of the system as a whole?

Table 2

Skill variations in four aspects of word problem solving in science and mathematics
(Adapted from Heller & Greeno, 1979)

	PROBLEM REPRESENTATION	KNOWLEDGE STRUCTURE	INITIAL ANALYSIS	SOLUTION STRATEGY
High Skill	Rely entirely on integrated mediating representation; semantic processing with qualitative elaboration; problem "understood"	Strong available schemata; solution procedures associated with schemata and prioritized for application	Use information associated with schemata to identify solution procedures; most promising procedure either immediately apparent or determined by preliminary traces.	Highly organized evokation of equations based on preliminary analysis and strong structural representation
Low Skill	Rely on verbal problem statement, syntactic processing	No evidence of relevant schemata; solution information not associated with problem category knowledge	Read problem only	Piecemeal, direct translation; "plug in values"; search for solution combined with equation manipulation

Table 3

A Delineation of Some Possible Kinds of Learning Errors

Individual Response Faults

Omission Errors
Intrusion Errors
Term Misuse
Misdefinition
"Sloppy " Errors and Idiosyncracies

Procedural Faults

Impasses
Production Deficiencies
Repair Patches
Mal-Rules
Premature Conclusions
Misgeneralizations and
Overgeneralizations
Inefficient Solution Paths
Incorrect Strategies
Inflexible Strategies
Faulty Monitoring or Checking

Declarative Faults

Chunk Omissions
Chunk Intrusions
Loose Couplings Within and Between Chunks
Misalignment of Chunks
Faulty Analogies
Inarticulations
Articulate Misconceptions
Natural Concept Misconceptions

Viewpoint Faults

Personal Biases
Misinterpretation of Content
Conflicting Personal Beliefs
Naive Theories

Figure Captions

Figure 1. A schematic topographic map leading to four related instructional goals.

Figure 2. Types of diagnosis and instructional steps indicated by reference task B4. Insert shows six possible routes from task B4 plus possibility of continued practice and coaching on that task.

Figure 3. Schematic three-dimensional views of domain topography showing:

a) two reference tasks with four measures for each;

b) an individual or class average profile at the end of a perfect course;

c) a class average profile suggesting sources of difficulty;

d) an individual mastering all beginner level tasks first;

e) an individual mastering all tasks in one region before entering other regions.

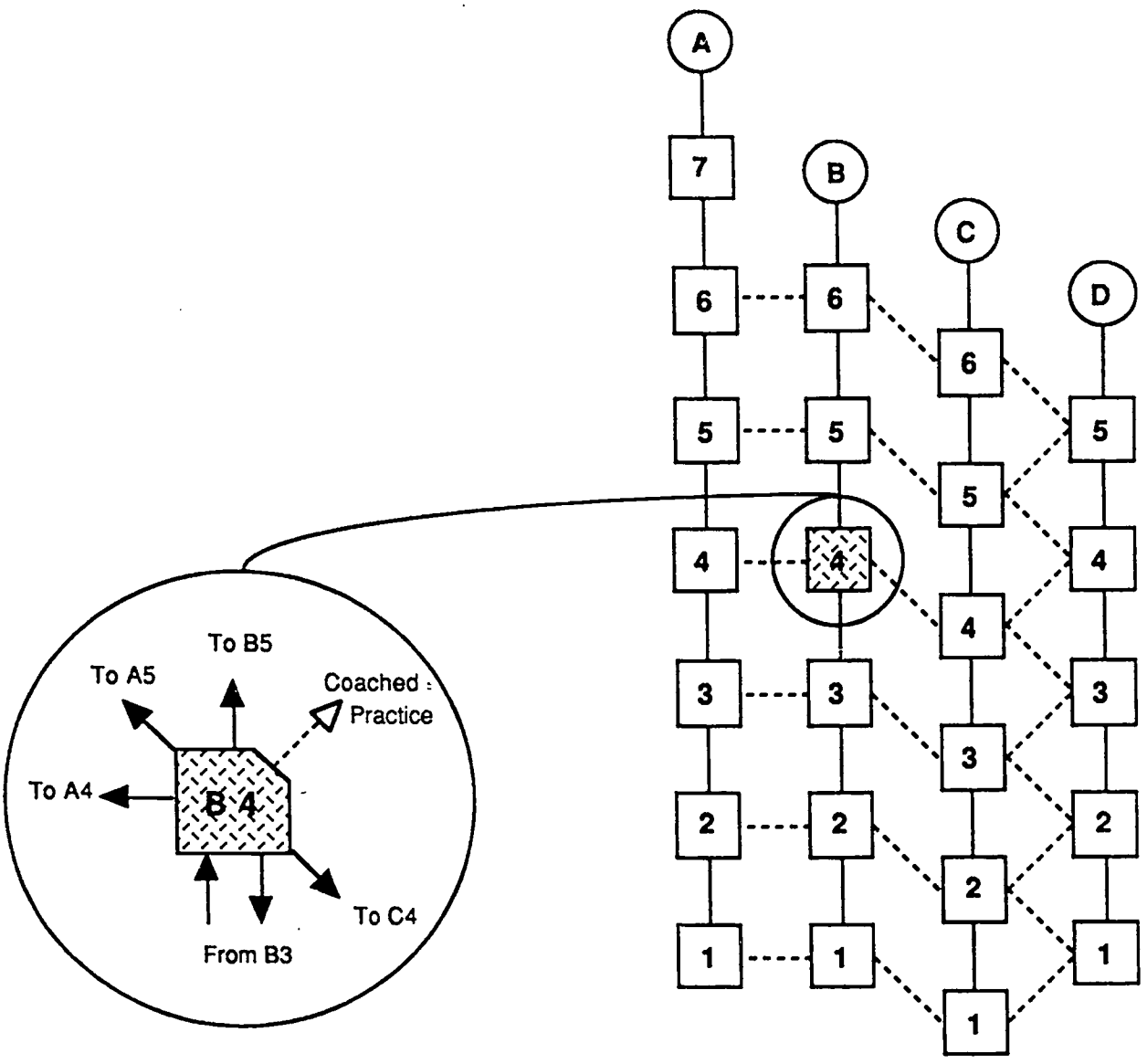
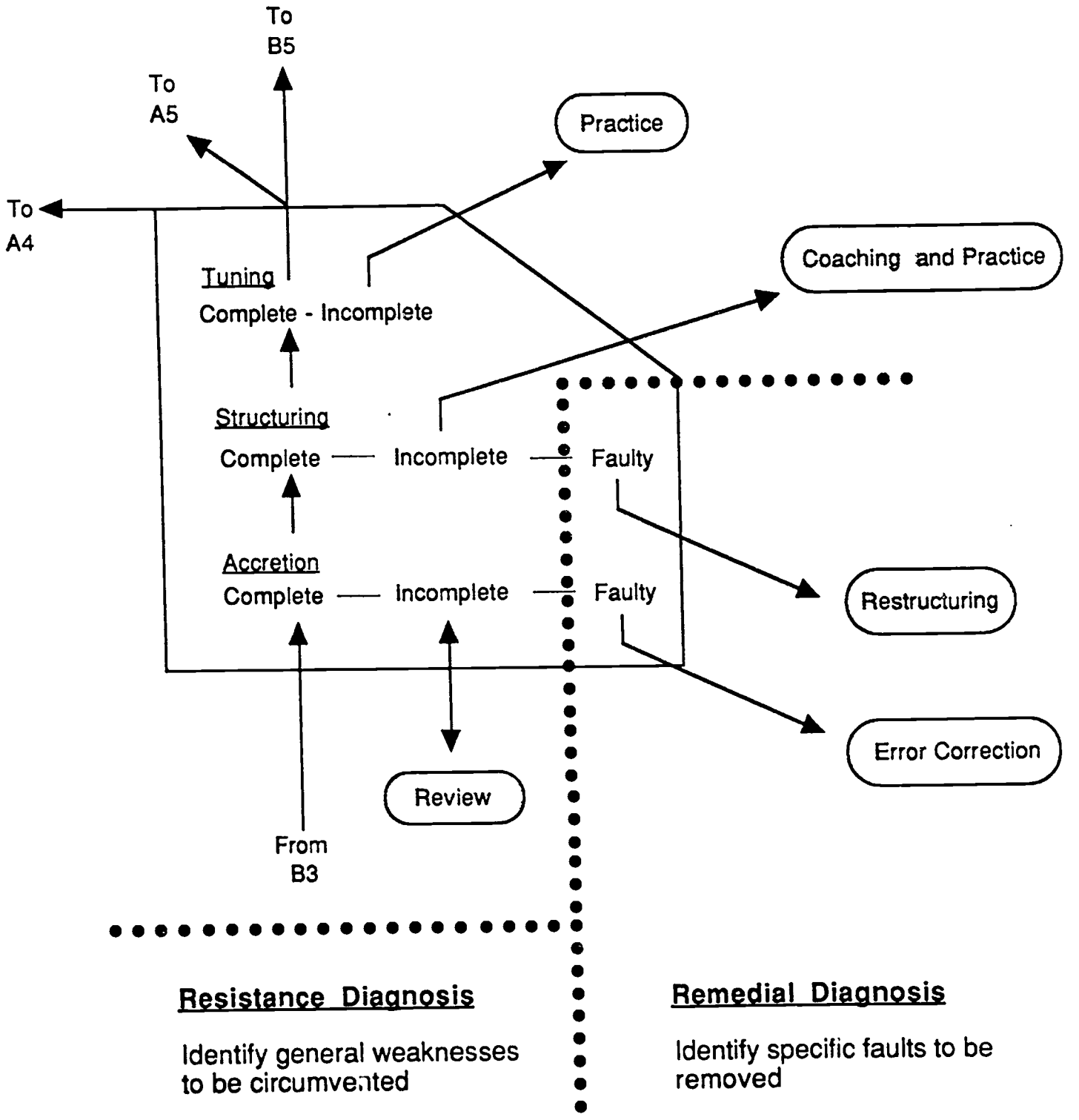

A
B
C
D
7
6
5
4
3
2
1
To B5
To A5
Coached
Practice
To A4
B 4
To C4
From B3

Figure 1.

Readiness Diagnosis

Choose nex. step that capitalizes on strengths and circumvents weaknesses

Identify needs for further instruction and practice



Resistance Diagnosis

Identify general weaknesses to be circumvented

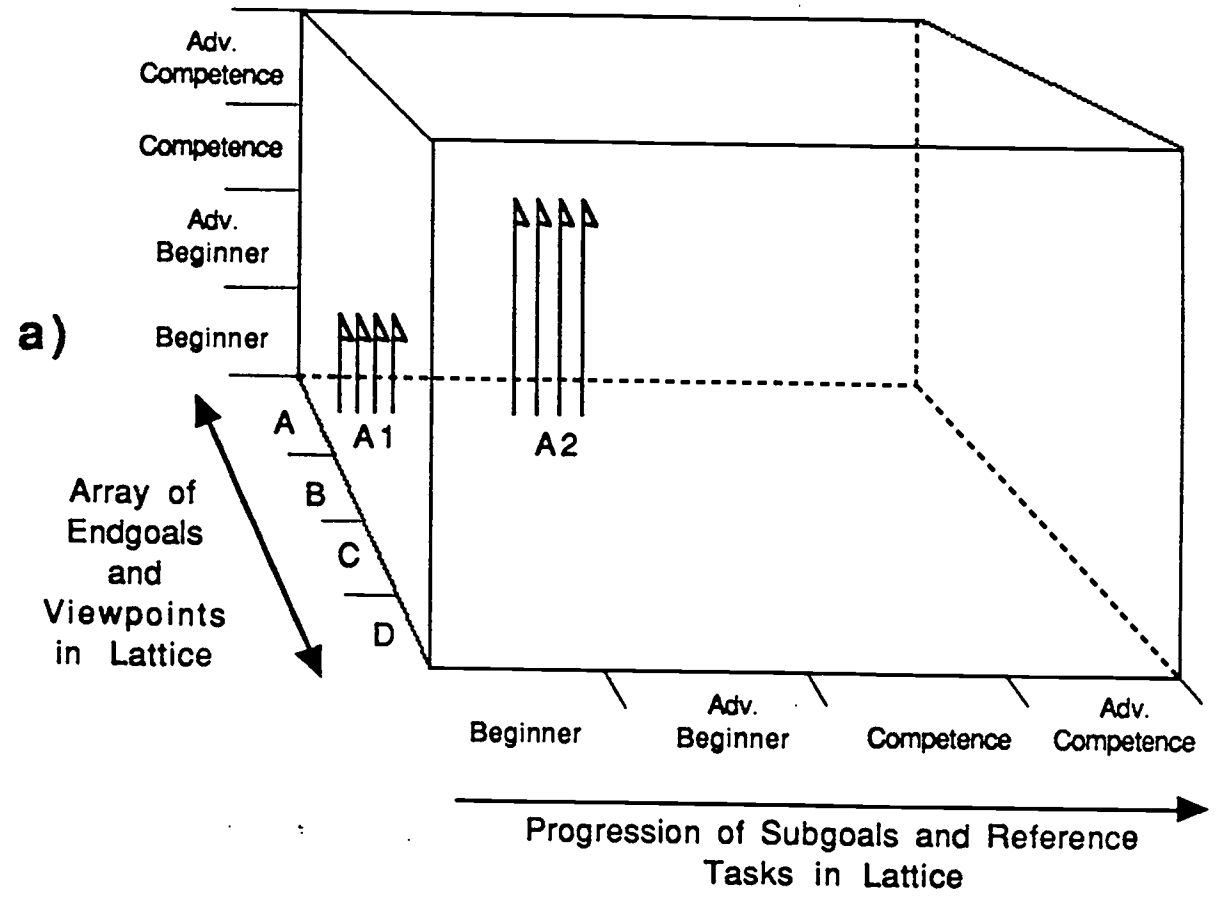
Remedial Diagnosis

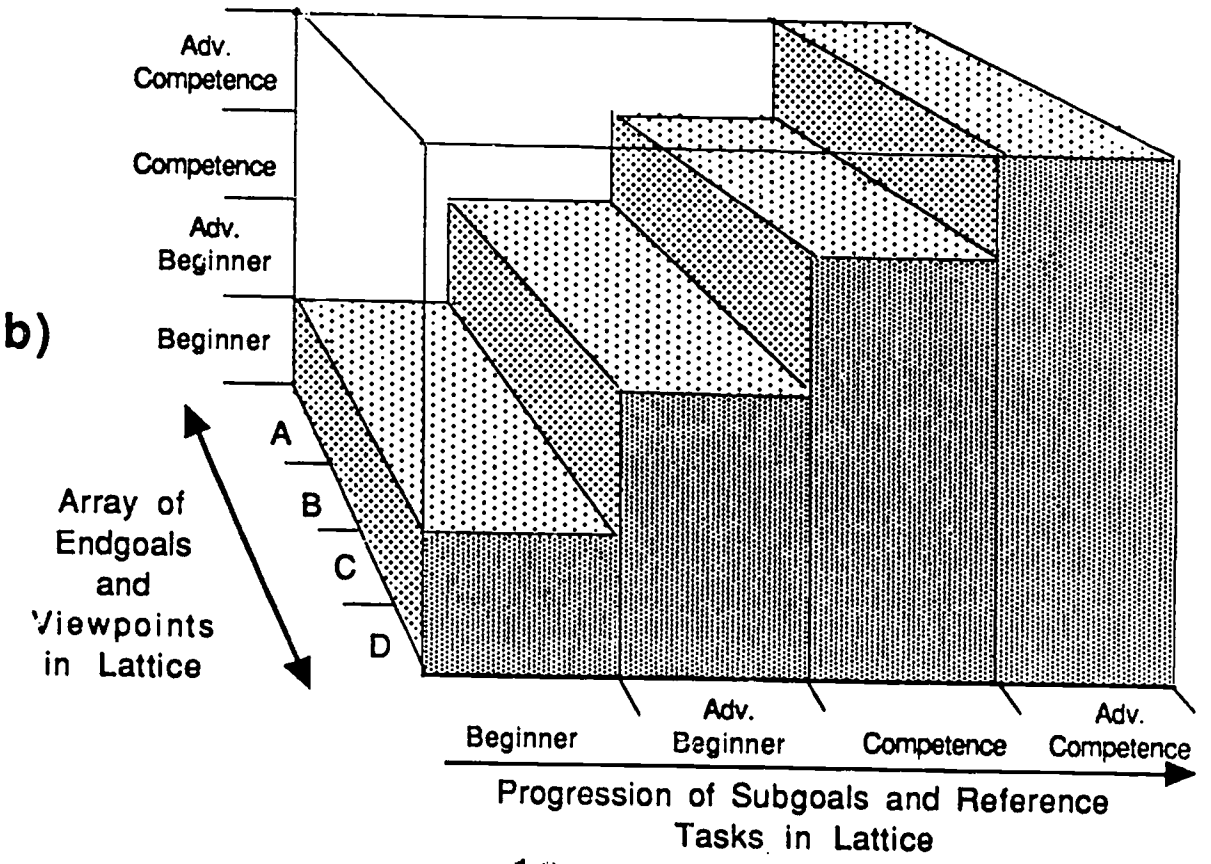
Identify specific faults to be removed

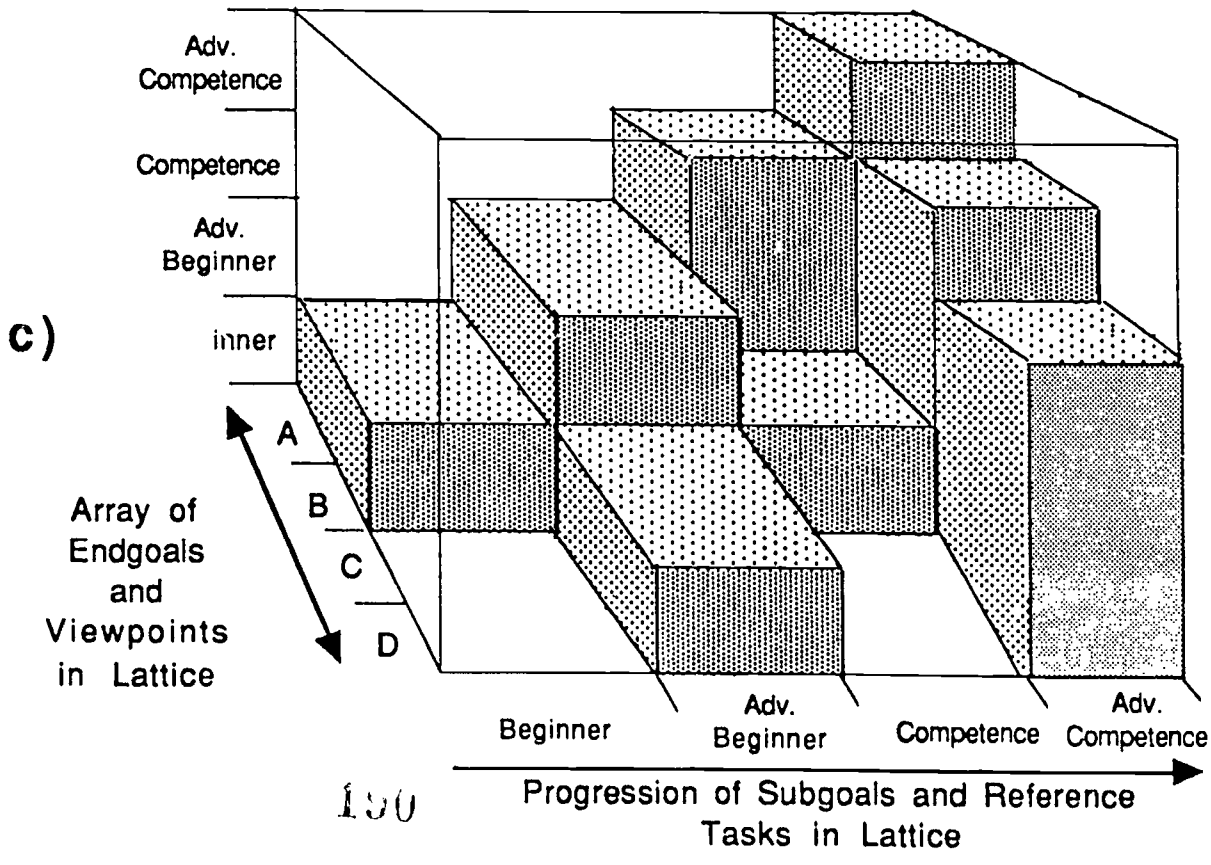
Figure 2.

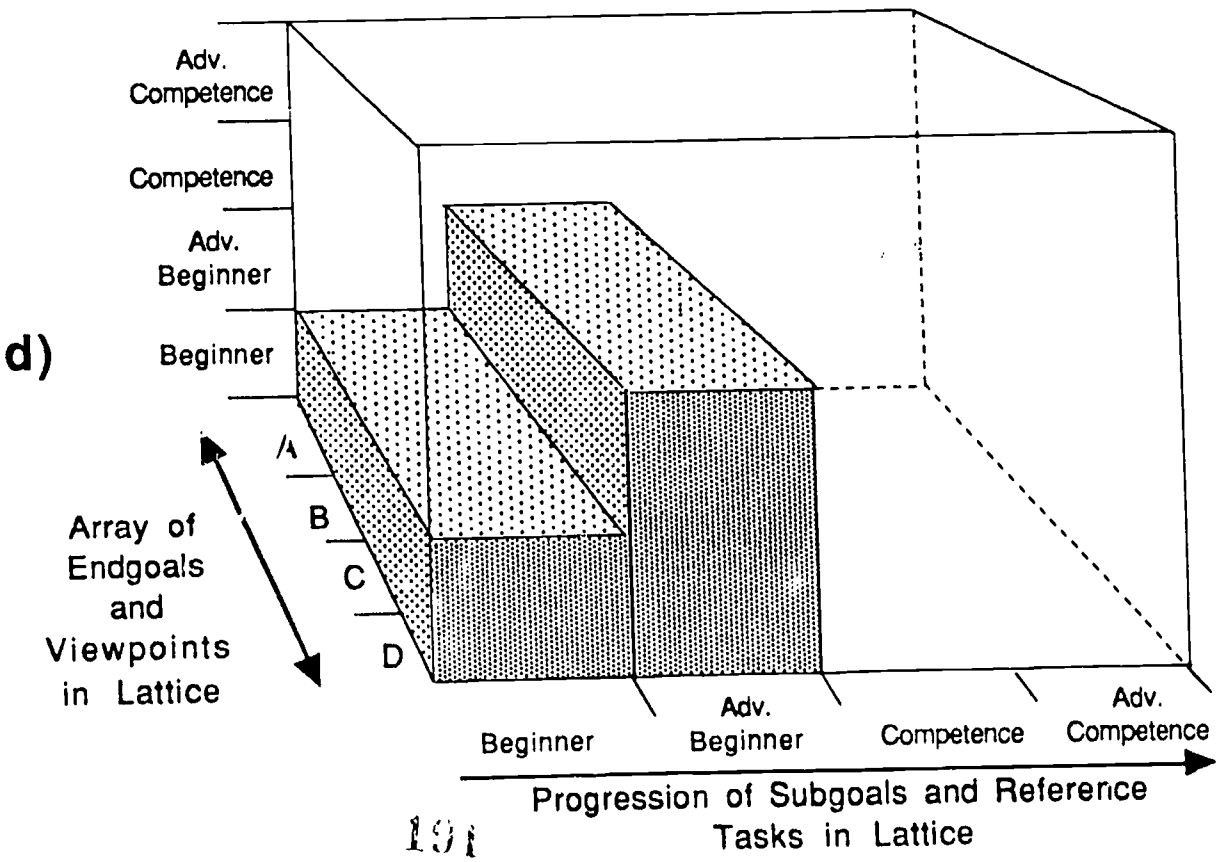
Figure 3.

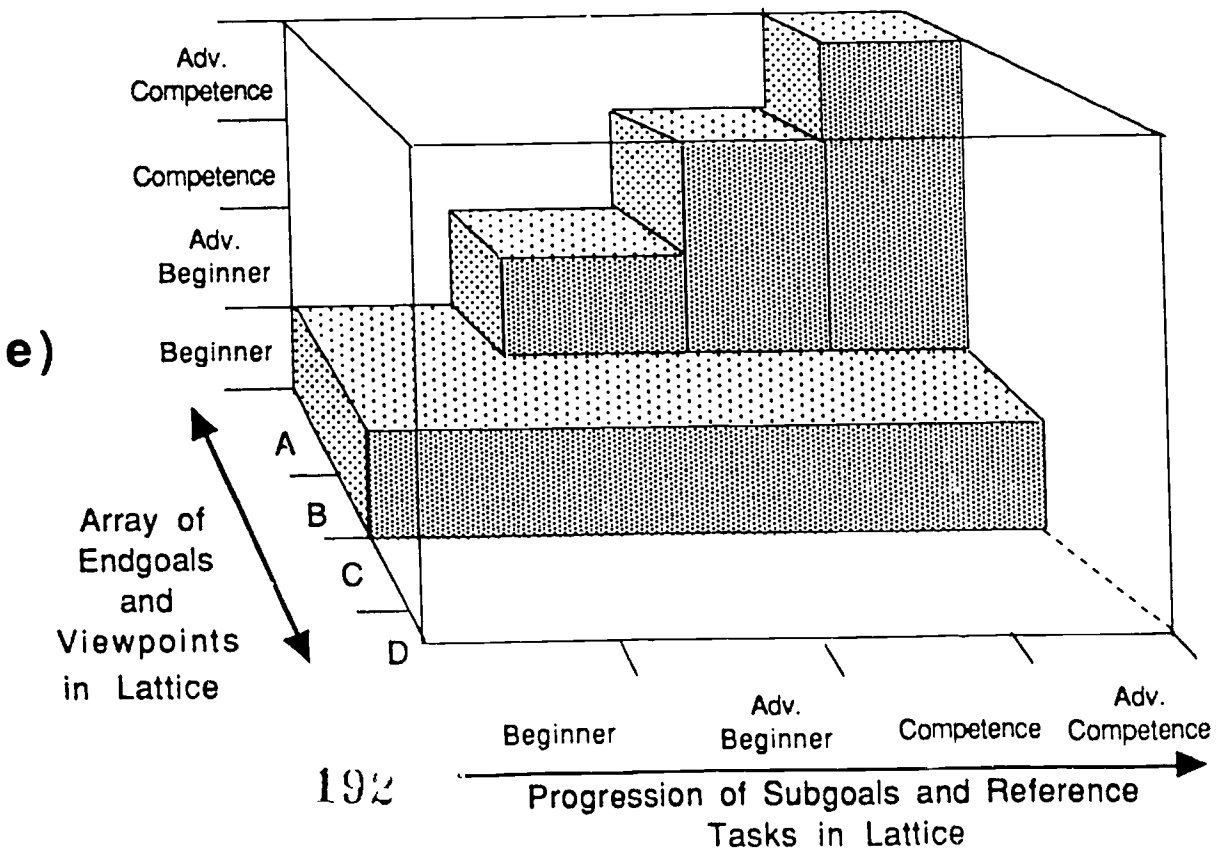
Schematic three-dimensional views of domain topography showing: a) two reference tasks with four measures for each; b) an individual or class average profile for a perfect course; c) a class average profile suggesting sources of difficulty; d) an individual mastering all beginner level tasks first; e) an individual mastering all tasks in one region before entering other regions.


a)
Adv. Competence
Competence
Adv. Beginner
Beginner
A
B
C
D
A1
A2
Array of Endgoals and Viewpoints in Lattice
Beginner
Adv. Beginner
Competence
Adv. Competence
Progression of Subgoals and Reference Tasks in Lattice


b)
Adv. Competence
Competence
Adv. Beginner
Beginner
A
B
C
D
Array of Endgoals and Viewpoints in Lattice
Beginner
Adv. Beginner
Competence
Adv. Competence
Progression of Subgoals and Reference Tasks in Lattice


c)
Adv. Competence
Competence
Adv. Beginner
inner
A
B
C
D
Array of Endgoals and Viewpoints in Lattice
Beginner
Adv. Beginner
Competence
Adv. Competence
Progression of Subgoals and Reference Tasks in Lattice


d)
Adv. Competence
Competence
Adv. Beginner
Beginner
A
B
C
D
Array of Endgoals and Viewpoints in Lattice
Beginner
Adv. Beginner
Competence
Adv. Competence
Progression of Subgoals and Reference Tasks in Lattice


e)
Adv. Competence
Competence
Adv. Beginner
Beginner
A
B
C
D
Array of Endgoals and Viewpoints in Lattice
Beginner
Adv. Beginner
Competence
Adv. Competence
Progression of Subgoals and Reference Tasks in Lattice