DOCUMENT RESUME

ED 361 405 TM 020 514

AUTHOR Nichols, Paul D.
TITLE A Framework for Developing Assessments That Aid Instructional Decisions.
INSTITUTION American Coll. Testing Program, Iowa City, Iowa.
REPORT NO ACT-RR-93-1
PUB DATE Apr 93
NOTE 46p.
AVAILABLE FROM ACT Research Report Series, P.O. Box 168, Iowa City, IA 52243.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Decision Making; *Diagnostic Tests; *Educational Assessment; Educational Objectives; Elementary Secondary Education; Higher Education; *Instructional Effectiveness; Program Evaluation; *Psychology; Research Methodology; *Test Construction
IDENTIFIERS *Psychology Referenced Tests

ABSTRACT

This paper is an attempt to organize the many loosely connected efforts to develop psychology-referenced assessments, focusing on the development of assessments to guide specific instructional decisions, sometimes referred to as diagnostic assessments. Many of the arguments apply to program evaluation as well, because assessments that reveal the mechanisms test takers use in responding to items or tasks provide important information on whether instruction is achieving its academic goals. Tests intended to select students for a particular educational institution of program are not considered. Societal trends that motivate the development of psychology-referenced assessment are outlined, and a framework is introduced within which the psychological and statistical aspects of an assessment can be coordinated. Efforts to develop psychology-referenced assessments are summarized in a five-step methodology that can guide future development efforts. Five tables and three figures illustrate the discussion. (SLD)

ACT Research Report Series 93-1

A Framework for Developing Assessments That Aid Instructional Decisions

Paul D. Nichols

April 1993

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

A FRAMEWORK FOR DEVELOPING ASSESSMENTS THAT AID INSTRUCTIONAL DECISIONS

Paul D. Nichols

ABSTRACT

In this paper, I attempt to organize the many loosely connected efforts to develop psychology-referenced assessments. I consider the development of assessments to guide specific instructional decisions, sometimes referred to as diagnostic assessments. Many of my arguments apply to program evaluation as well--assessments that reveal the mechanisms test takers use in responding to items or tasks provide important information on whether instruction is achieving its academic goals. In contrast, I do not address tests intended to select students for a particular educational institution or program even though I believe a psychology-referenced approach could benefit the development of such tests, as well. My goals in this paper are 1) to outline the societal trends that motivate the development of psychology-referenced assessment, 2) to introduce a framework within which the psychological and statistical aspects of an assessment can be coordinated, and 3) to summarize efforts to develop psychology-referenced assessments in a five-step methodology that can guide future development efforts.

ACKNOWLEDGMENTS

I am grateful to Robert L. Brennan, Dean Colton, David F. Lohman and David J. Mittelholtz for their helpful comments on drafts of this article.

A FRAMEWORK FOR DEVELOPING

ASSESSMENTS THAT AID INSTRUCTIONAL DECISIONS

Several years ago, I examined the responses of students to figural analogies from a widely used, nationally standardized test battery. I intended to explain test takers' solutions to analogies by fitting process models of analogical reasoning to the responses of individuals. As I attempted to code the items for analysis, I realized I could not explain test takers' solutions because the figural analogies were not constructed to reveal the cognitive mechanisms test takers used to respond. Efficient diagnosis of test takers solutions required items or tasks constructed systematically to reveal differences in those solutions. Subsequently, I constructed my own set of figural analogies based on a model of how test takers might solve analogies (see Nichols, 1990; Nichols, 1991). The experience convinced me that researchers following a new approach combining cognitive science and psychometrics could successfully develop assessments that reveal mechanisms test takers use in responding. These new assessments could inform instruction directly by describing the students' processes and knowledge structures that are the targets of instruction.

Over the past decade or so, a growing number of writers have argued that cognitive science and psychometrics could be combined in the service of instruction (Bejar, 1984; Haertel & Calfee, 1983; Linn, 1986; Messick, 1984; Snow & Lohman, 1989). They criticized traditional testing for losing sight of the psychology of the performance being tested (Glass, 1986; Glaser, 1981). Traditional testing practices appear to place more emphasis on statistical technique than on the psychology of the construct being measured (Anastasi, 1967). Given some knowledge of the goals and methods of instruction and of the psychology of the construct, "educational tests might be made more diagnostic of malfunctions in learning and more informative for instructional adaptation" (Snow & Lohman, 1989, pp. 266).

Researchers have progressed beyond what Pellegrino (1992) has called verbal statements of intent and handwaving about proposed solutions to the hands-on business of researching and developing diagnostic assessments combining cognitive science and psychometrics, what I call psychology-referenced assessment (PRA)[1]. They are constructing new assessments informed by research on the psychology of learning and achievement and embracing new statistical models for combining observations. They design

problems and tasks through systematic, research-based variations in problem characteristics. The defining characteristic of PRA is that it makes explicit the substantive assumptions the test developer is using to construct test materials and assign scores. These substantive assumptions describe the knowledge and skills a performer in the test domain would use, how the knowledge and skills develop and how more competent performers differ from less competent performers. These substantive assumptions are testable because they are explicit.

PRA is not alone in emphasizing the mechanisms students use in responding to items or tasks or the use of assessment results to inform instruction directly. Student modeling shares an interest with PRA in representing the cognitive processes and structures of learners and using such information to make pedagogical decisions. A student model is a representation within an intelligent tutoring system (ITS) of the student's processes and knowledge structures that have repercussions for learning. The challenge in ITS is to form and update a student model of moment-to-moment changes in processes and knowledge. Certainly, PRA could be used to develop student modeling approaches. Elements of PRA are similar to some elements of ITS. For example, the substantive assumptions of PRA include knowledge the student is expected to acquire, a part of the ITS domain knowledge component, and how knowledge is acquired, an aspect of the ITS pedagogical knowledge component. Furthermore, the representation of the student that is the result of PRA may be conceived as a subset of an extended student model. However, PRA is distinguished from student modeling by the grain size of the behavior that is the focus of inference. For student modeling, the focus is on detailed diagnosis of moment-to-moment changes relevant to the didactic decisions required for the tutorial interchange (Wenger, 1987). In contrast, the focus of PRA may be on grosser changes relevant to broader decisions such as placement.

Traditional diagnostic tests share an interest with PRA in informing instruction. According to Anastasi (1982), "Diagnostic tests are designed to analyze the individual's specific strengths and weaknesses in a subject and to suggest causes of his or her difficulties" (p. 415). There are currently a large number of diagnostic tests including the Nelson Denny Reading Test, the Stanford Diagnostic Mathematics Test, the Stanford Diagnostic Reading Test, and the Instructional Tests of the Metropolitan

Achievement Tests. These tests differ from PRA assessments in three aspects: 1) the design is based on logical taxonomies and content specifications and lacks explicit psychological models of the structures and processes that underlie domain performance (Snow & Mandinach, 1989); 2) the scores are tied to content areas rather than cognitive mechanisms; and 3), the scores are computed often using methods developed to select students most likely to succeed in a uniform instructional environment rather than methods developed to make inferences about cognitive structures and processes.

In this paper, I attempt to organize the many loosely connected efforts to develop psychology-referenced assessments. I consider the development of assessments to guide specific instructional decisions, sometimes referred to as diagnostic assessment. Many of my arguments apply to program evaluation as well--assessments that reveal the mechanisms test takers use in responding to items or tasks provide important information on whether instruction is achieving its academic goals. In contrast, I do not address tests intended to select students for a particular educational institution or program even though I believe a psychology-referenced approach could benefit the development of such tests, as well. My goals in this paper are 1) to outline the societal trends that motivate the development of psychology-referenced assessment, 2) to introduce a framework within which the psychological and statistical aspects of an assessment can be coordinated, and 3) to summarize efforts to develop psychology-referenced assessments in a five-step methodology that can guide future development efforts.

Societal Demand for New Assessment Techniques

When I was examining the responses of students to figural analogy test items, I intended to explain to some degree differences in test scores through differences in test takers' use of analogy. Several problems plague this pursuit (see Lohman, in press), but I was frustrated because tests are not traditionally constructed to elicit qualitative differences in strategy. Other researchers, also frustrated by the lack of diagnosticity of traditional tests, have designed assessments to reveal the mechanisms used by tests takers in responding. In this section, I sketch the societal trends that have motivated, at least in part, development of diagnostic assessments. I will not pretend to be an observer of social trends, but I will review what others have observed. I present this context to highlight the differences in purpose between PRA and

traditional tests and to describe the role PRA may assume in education.

Traditional assessments were developed to confront the dilemma of educators in the early part of the 20th century. Educators were faced with determining which students would be able to profit best from uniform instruction designed essentially for the majority of the population (Glaser, 1981). Resources limited the information they could gather about each student and precluded tailoring programs to individual students' needs (Mislevy, in press). I largely agree with Mislevy (in press) who argues that current test theory has been effective in selecting students most likely to succeed in a particular educational institution or program.

Traditional assessments developed to identify which students would profit best from a uniform instructional environment are based generally on what Snow and Lohman (1989) have termed the educational psychometric measurement (EPM) approach. The test theories that dominate the EPM conception are aimed at estimating a person's location on an underlying latent variable--a true score in classical test theory (CTT) or a latent trait in unidimensional Item Response Theory (IRT). This location is typically interpreted as an amount on the latent scale. The model is judged as to how well it places people into a single sequence or aids selection into a single program (Mislevy, in press). Either CTT or IRT may usefully inform decisions about such linearly ordered alternatives (Dawes & Corrigan, 1974).

Work on PRA has been motivated, at least in part, by the current emphasis on helping individuals to succeed in educational opportunities in contrast to selecting individuals for those opportunities (Stiggins, 1991). The requirement now is to design education that helps all students succeed (Carnegie Council on Adolescent Development, 1989; National Education Goals Panel, 1991; National Governors' Association, 1990). According to the National Governors' Association Report on the Task Force on Education, "We must abandon the view that only a small proportion of our population must be well educated, that many can get by with less knowledge and fewer skills" (pp. 7). A source of the current concern for the learning needs of all children is the recognition that there is a strong connection between how well a nation can perform and the existence of high-quality, widely distributed education (American Association for the Advancement of Science, 1989). Educators and policy makers are demanding new assessments to help

individuals succeed in educational opportunities; they require assessments to evaluate school learning and inform directly instruction.

The EPM approach toward constructing assessments may not be helpful for designing assessments used to evaluate school learning and inform directly instructional decisions. As Bejar (1984) notes, scores derived from traditional CTT or IRT approaches provide only general information to guide specific instructional decisions. "Thus, the student with a lower score could benefit, perhaps, from additional or remedial instruction, but there are no guidelines for developing efficient courses of instruction" (pp. 185). Scores on new performance-based or authentic assessments often provide little more information than traditional assessments to guide specific instructional decisions. Performance-based or authentic assessments may well consist of tasks that are more representative of some intended domain. However, scores on these assessments, often holistic ratings, provide little more information than how well one student compares to another or how a student's performance compares to a criterion value. In either case, scores indicate no more than the need for additional instruction.

In response to the demands for new measures that help individuals succeed in educational opportunities, researchers have constructed assessments informed by research on the psychology of learning and achievement and embracing statistical models for making inferences regarding the structures and processes that underlie domain performance. These assessments have been developed using a PRA approach in contrast to the EPM approach used in traditional assessments. The PRA approach can directly inform instructional decisions by focusing on the knowledge and processes that are the instructional targets for educational reform programs such as Project 2061 (Amerian Association for the Advancement of Science, 1989). By describing the processes and knowledge structures students are expected to acquire, test developers are able to construct a task or an item that is predicted to demand those processes and structures. Conversely, test developers are able to infer the processes and knowledge structures used by an individual to respond to an item or task. Thus, the PRA approach is distinct from traditional approaches relying on logical taxonomies and content specifications and employing statistical approaches developed for selecting students.

Framework for Psychology-Referenced Assessment

After deciding to construct a set of figural analogies to identify test takers' solution strategies, I realized that I was lacking a test theory that allowed me to diagnose their strategies This required more than summing the correct responses[2]. I was searching for a new test theory-- a test theory suited to diagnosing learners use of analogy. In this section, I attempt to describe the psychological and statistical considerations involved in adopting a test theory This may seem a strange idea to many measurement specialists and psychologists, alike. I argued in the previous section that decisions regarding the nature of instruction require a different approach toward assessment than decisions regarding selection for uniform instruction. In this section, I argue that different conceptions of domain performance require different test theories All domains are not alike, and the test developer should not ignore the unique nature of the performance. I use examples from subtraction to illustrate this argument because subtraction has been studied extensively by psychologists and has, perhaps not coincidently, been the focus of a number of PRA research efforts.

Following Ippel (1986, 1991) and Lohman and Ippel (in press), I view test theory as consisting of two related aspects: an observation design for constructing and arranging observations and a measurement design for collecting and combining responses. Current test theories confound these two aspects. The observation and measurement designs provide the test developer a framework within which the psychological and statistical considerations can be coordinated. I will argue later in this section that the validity of both designs must be evaluated within the context of substantive research. The observation and measurement designs are discussed separately but practical issues concerning the observation and measurement designs should be considered in company because diagnosis is possible only through their coordination. The most sophisticated inference from any diagnostic measurement design is limited by the richness of the performance elicited through the observation design. Alternatively, the value of performance elicited through the observation design is limited by the measurement design's power to use the information.

Observation design.

The observation design describes the characteristics of assessment tasks or items that make demands on the test taker, how these characteristics are to be organized in the construction and ordering of observations. and the nature of the responses required. The purpose of the observation design is to construct and arrange observations in a way that reveals the mechanisms test takers use in responding. This implies the specification of the demands made on the test taker by the assessment tasks. Any such specification of task demands requires substantive research on the psychology of the test domain. The substantive research in test development identifies what the task or item characteristics are so that they can be systematically manipulated to investigate the cognitive processes and structures influenced by each. Typically. test developers are not prepared to identify or manipulate these characteristics (Snow & Lohman, 1989).

An example of manipulating task characteristics to investigate the mechanisms used by test takers is provided by Tatsuoka (1990) for mixed fraction subtraction. A pair of fraction subtraction items is shown in Figure 1. A seventh- or eighth-grader may mistakenly believe that it is necessary to reduce the whole number by 1 and add 10 to the numerator of the first fraction when problems, such as the items in Figure 1, require the student to increase the numerator of the first fraction. This misconception will produce the correct answer if the denominator of the first fraction happens to be 10. In order to detect the misconception, one item must have a denominator not equal to 10 in the first fraction and another item must have the denominator equal to 10. The pair of items in Figure 1 would discriminate this misconception from the correct procedure for fraction subtraction.

Another example of manipulating task characteristics to investigate the mechanisms test takers use in responding is provided by VanLehn (1982). The test shown in Table 1 was developed to diagnose the buggy performance of third and fourth graders doing multicolumn subtraction. This test is designed so that students using any of at least 43 bugs would miss two or more problems. Every possible diagnosis was distinguished from the others by the response on at least one problem. Thus, a student who systematically follows the Subtract Smaller From Larger rule never borrows but subtracts the smaller digit from the larger

digit in each column. The student would produce the pattern of responses shown for Rule A of Table 1 In contrast, a student systematically following the Stops Borrow at Zero rule adds 10 to the current column but does not decrement from the column to the left if that column is a zero. That student would produce the other pattern of responses shown for Rule B of Table 1. As these two examples illustrate, item characteristics can be systematically manipulated to reveal the mechanisms test takers use in responding.

Insert Table 1 about here

As these examples illustrate, the validity of the observation design is based on substantive research in the test domain because the characteristics that make demands on the test taker are identified through substantive research. This underscores the importance of psychology in test development. For example, the multicolumn subtraction test reported by VanLehn (1982) was developed only after extensive research on the buggy procedures of grade school students. A database of bugs was developed after reviewing subtraction solutions from more than two thousand students (Brown & Burton, 1978). Items were systematically constructed so that a set of items could discriminate between students who may have a number of different known subtraction bugs. This approach is fundamentally different from a test developer following a content sampling approach and constructing a certain number of 1-column, 2-column or 3-column subtraction problems.

Measurement design.

The measurement design defines the object of measurement and describes the procedure or set of procedures to assign a value or category to an object of measurement. In addition, the measurement design must provide ways of addressing the precision of the procedure for assigning a value or category. Test takers make careless mistakes responding to tasks or items and the measurement design must account for this when expressing precision associated with the assignment of a value. Finally, the measurement design must define the diagnostic value of classes of tasks or items defined by characteristics identified in the observation design.

For example, the ACM system of Langley, Wogulis, and Ohlsson (1990) uses artificial intelligence (AI) and statistical methods to generate a production system model of a student's performance on multicolumn subtraction problems. In this measurement design, the object of measurement is the student and the set of procedures consists of the AI and statistical methods used to assign a production system model to the student. A production system is a set of condition-operation pairs (Anderson, 1983). The condition identifies certain data patterns and the operation executes if the data patterns match elements in working memory. The system begins with a description of a set of subtraction problems, a set of responses from a student, and a set of mental operations the student may have applied to solve the subtraction problems. An example of a set of operations for subtraction are shown in Table 2. For each problem, the system generates a set of solution paths, called path hypotheses, a student may have followed to respond to a problem. The system produces a production system that identifies inappropriate conditions under which that student may have applied the subtraction operations.

Insert Table 2 about here

The ACM system uses an AI procedure called *heuristic search* to generate a production system for a student on a particular subtraction problem. Using heuristic search, the mental operations identified earlier are applied to the initial description of the subtraction problem to produce an intermediate state along the solution path. The operations are applied to each successive intermediate state until the student's response to the problem is generated.

The ACM system uses the X^2 statistical procedure to find the production system that most nearly describes the student's responses across the set of subtraction problems. The X^2 statistic is used with an AI procedure called *learning from solution paths* (Sleeman, Langley, & Mitchell, 1982). The system produces positive and negative applications of an operation. A positive application of an operation generates a state on a possible solution path; a negative application of an operation generates a state off a possible solution path. The system adds conditions to operations so that the operation would apply in

the most positive applications and the fewest negative applications of the operation. The conditions on each operation allow the production system to follow the solution path and avoid paths that do not lead to the student's response. The ACM system employs a forward approach in selecting the conditions to add to an operator. The system considers individual conditions in turn and selects that condition that, when added to the action to form a production rule, produces the greatest X^2 value. This process continues until enough conditions have been added to the action so that none of the negative applications are produced. Then the system begins a backward approach of dropping each condition from the production rule unless dropping that condition from the production rule significantly reduces the X^2 value for that rule. The newly constrained production system is applied to the next problem and the system begins searching for possible solution paths.

The validity of the measurement design is supported by substantive research suggesting appropriate procedures for combining observations to assign a value to an object of measurement. For example, the decision model in the ACM system is consistent with the system developers' psychological assumptions regarding students' subtraction problem solving. First, the system generates a set of possible solution paths because the system developers assumed that all problem solving involves search through some problem space (Langley, Wogulis, & Ohlsson, 1990; Ohlsson, 1990). Second, the system produces a production system that identifies inappropriate conditions under which that student may have applied the subtraction operations because the system developers assumed that students' subtraction errors are due to rules with the correct actions but the incorrect conditions. Third, the system uses the X^2 statistic to discriminate between competing conditions because the system developers assumed students slip sometimes when responding and so obtained response patterns would not match exactly frequently predicted response patterns.

Generalizability theory.

Generalizability theory (GT) provides a conceptual framework and a set of statistical procedures for addressing a broad scope of measurement issues (Cronbach, et. al., 1972; Brennan, 1992). I present GT as an extension of classical test theory. GT has both an observation design and a measurement

design. The observation design I associate with GT describes the set of measurement conditions, or universe of generalization, to which generalization of the test score is intended. The universe is characterized by facets, such as items and occasions, that may be crossed or nested. This description necessitates the specification of what conditions of measurement contribute to error. Given that a universe of generalization has been identified, a measurement procedure is viewed as consisting of a random sample of conditions of the facets. Efforts have been made to address issues of representation. For example, items may be constructed according to specifications that represent the distribution of content categories across major textbooks. However, efforts to represent content are only vaguely directed at revealing mechanisms test takers use in responding to items or tasks.

The measurement design I associate with GT involves estimating a persons' universe score. A person's universe score can be conceptualized as an average (or expected value) observed score over all randomly parallel measurement procedures associated with the universe of generalization. Under this measurement design, the object of measurement is the person. The universe score, like the true score in CTT, is an average score. Each item contributes equally to the universe score and the order of averaging or summing items is arbitrary. Sources of error are represented as facets in the universe of generalization. GT clearly differentiates among multiple sources of random error that are undifferentiated in CTT. GT applies certain analysis of variance procedures as a way of differentiating among multiple sources of error in persons' observed scores. A substantial literature has been established addressing the precision of estimating universe scores (see Feldt & Brennan, 1989)

A comparison of the Stanford Diagnostic Mathematics Test (SDMT; Beatty, Madden, Gardner, & Karlsen, 1976) for grades 6 and 7 with VanLehn's (1982) diagnostic test illustrates differences between assessments constructed using classical test theory and assessments constructed using the PRA approach. Within the computation subtest, the SDMT has a concept/skill domain labeled Subtraction of Whole Numbers. The observation design for the Subtraction of Whole Numbers concept/skill domain is represented by the item objectives. The objective for each item in the Subtraction of Whole Numbers concept/skill domain is shown in Table 3. As Table 3 shows, the focus of the item objectives is on item

content--the value of the number (tens, hundreds, thousands) or the location of the zero (tens or hundreds place) rather than the subtraction processes used by the student. In contrast, VanLehn's test focuses on the students' use of procedures such as borrow, decrement and carry. Of course, item objectives for the Subtraction of Whole Numbers concept/skill domain may be written to focus on procedures such as borrow, decrement and carry. However, the usefulness of information elicited through any such revised observation design is limited by the measurement design for the Subtraction of Whole Numbers concept/skill domain

Insert Table 3 about here

The p X I design in GT (Brennan, 1992) perhaps describes best the measurement design for the Subtraction of Whole Numbers concept/skill domain. A value is assigned a person by computing the average correct or number correct. Interaction between people and items is treated as error under the p X I design. In contrast, VanLehn's (1982) diagnostic test uses that same information (i. e., different response patterns) to diagnose buggy performance.

The measurement designs of PRA and GT (or CTT) treat differently the information in the matrix defined by persons and items. PRA uses the information from the interaction between persons and items in the matrix--test takers different response patterns. On VanLehn's (1982) diagnostic test, a test taker who systematically follows the Subtract Smaller From Larger rule would produce a different pattern of responses and a different diagnosis than a test taker systematically following the Stops Borrow at Zero rule. GT uses the information in the marginals of the matrix. On the SDMT's Subtraction of Whole Numbers subtest, test takers differ in the number correct. These two sources of information from test takers' responses appear not to overlap. Thus, test takers assigned different diagnostic categories may be assigned the same number correct score. This difference between the measurement designs of PRA and GT makes the two approaches irreconcilable.

In summary, current assessments using GT are based on an observation design that describes the universe to which generalization of the test score is intended. Generally, the universe of generalization is

characterized by descriptions of content. The measurement design of GT addresses the sources of error in averaging or summing a test taker's performance over items. The focus of the GT measurement design on average scores is irreconcilable with the focus of the PRA measurement model on patterns of responses. In contrast, the observation design of GT provides a valuable conceptual framework for understanding the limits of generalizing from performance on a particular diagnostic assessment to a broader educational setting. Test developers of PRA should explore modifying the observation design of GT so as to characterize the universe of generalization by the mechanisms test takers use in responding.

Generally, assessments constructed using GT (or CTT) have no explicit substantive model of the psychology of the test domain. However, a mouse trap is difficult to build without some ideas about mice; An achievement test is difficult to build without some ideas about achievement. According to Mislevy (in press): "Standard test theory evolved as the application of statistical theory with a simple model of ability that suits the decision-making environment of most mass educational systems" (pp. 1). The substantive theory appears to be implicit in CTT and reflects the assumptions of psychological theories of the early 20th century (esnick & Resnick, 1992). According to these early theories, differences in competence are assumed to be due to differences in the accumulation of facts and skills. These psychological assumptions do not reflect necessarily the measurement specialist's level of sophistication in psychology, although Shepard (1991) reports that many measurement specialists hold views of learning consistent with these early psychological theories. However, as Anastasi (1967) lamented:

> Psychometricians appear to shed much of their psychological knowledge as they concentrate upon the minutiae of elegant statistical techniques. Moreover, when other types of psychologists use standardized tests in their work, they too show a tendency to slip down several notches in psychological sophistication (pp. 300).

Using tests constructed following CTT, differential psychologists proposed theories that described the organization of individual differences in terms of traits. The conception of individual differences in terms of static traits is tied closely to the EPM test theory. Individuals differ in the amount of a trait possessed. Using factor analysis, researchers identified traits by how well assessments from different content areas

ranked individuals differently The conception of individual differences in terms of static traits made impossible a better understanding of the mental processes of performers in a domain (Lohman & Ippel, in press). As McNemar (1964) observed with regards to intelligence: "Indeed, it is difficult to see how the available individual difference data can be used even as a starting point for generating a theory as to the process nature of general intelligence," (pp. 8).

In summary, current assessments using GT (or CTT) are based apparently on substantive assumptions regarding learning that are implicit in the test theory. Test theory may be compared to data, as when Lord (1980) exhorts, "Such mathematical models can be used with confidence only after repeated and extensive checking of their applicability" (pp. 15). But these implicit substantive assumptions are never checked against competing assumptions and so no measure of the adequacy of fit is available. "Just how poor a fit to the data can be tolerated cannot be stated exactly because exact sampling variances are not known" (Lord, 1980, pp. 15). One purpose of the test development approach introduced in the next section is to make explicit the role of psychological assumptions in the test development process.

Methodology for Psychology-Referenced Assessment

While trying to construct a diagnostic assessment of learners use of analogy, I reviewed the diagnostic assessment work of other researchers. In their work, I perceived a general pattern that I subsequently refined into a series of steps that constitute a methodology for developing PR assessments. I do not conceive these steps as discrete stages nor do I conceive the sequence as inviolable. The description of steps is simply a useful device to communicate the activities associated with developing PR assessments.

In this section, I describe a methodology for developing PRA within which is coordinated the substantive and statistical aspects of test theory. The five steps in this methodology for developing a psychology-driven assessment are presented in Table 4. Each step is described briefly. To illustrate these steps, I recount the development of Gitomer's diagnostic logic gate assessment[3] as I have followed it through publications and presentations. I use Gitomer's research to illustrate this methodology rather than my own work on analogy because his work has matured to produce a viable assessment whereas my

efforts are preliminary. I emphasize that Gitomer has never claimed to have followed these steps.

Insert Table 4 about here

Substantive base.

The first step in this methodology and the foundation of psychology-referenced test development is the construction of a substantive base. A substantive base is constructed from original research and research reviews but also includes assumptions about how to represent best learning and individual differences. A useful substantive base includes two elements (see Figure 2): (1) a description of the knowledge and skills a performer in the test domain would use. The description may include how the knowledge and skills develop and how more competent performers differ from less competent performers: (2) a specification of task or item characteristics that are hypothesized to influence the knowledge and skills used by performers in the domain. Taken together, a description of the performer's knowledge and skills and a specification of task or item characteristics constitute the construct representation of an assessment (Embretson, 1983). The substantive base is a dynamic element in test development; substantive research by the test developer and by others continues during and after the assessment is developed.

The substantive base is consulted in every stage of test development. As I argued when describing the observation and measurement design, the substantive base provides the rationale for both the observation and the measurement design. Furthermore, the substantive base indicates limits in the generalization from scores. The importance of the substantive base was described well by Messick (1989) who referred to it as construct theory:

> Construct theory as a guide to test construction provides a rational basis for selecting task content, for expecting certain consistencies in item responses, and for predicting score relationships. If one starts with well-grounded theory, the whole enterprise is largely deductive in nature, and the approach to construct validation can be well specified and relatively rigorous (pp. 49).

The substantive base enables test developers to create a set of items or tasks and infer the knowledge and skills used to respond to those tasks. The premise is simple: by recognizing the different structures and skills that an individual brings to a task or test, the test developer should be able to construct a task or an item that requires those structures. Conversely, the test developer should be able to infer the knowledge and structures used by an individual to respond to an item constructed to reveal such knowledge and structures. The necessary conditions are that the test developer understands the structures and skills required by performers on each task and that some subset of responses will discriminate individuals who differ on one or more of those skills (Gitomer & Yamamoto, 1991).

As the work of Gitomer and his colleagues illustrates, the construction of a substantive base to support test development is a laborious task. Gitomer's early work in this area focused on defining those skills which characterized competent performance of avionics technicians (Gitomer, 1984) The avionics technician is asked to identify and repair malfunctions in airborne avionics equipment and maintain the troubleshooting equipment. Skills that differentiate more competent from less competent technicians were explored using two approaches. First, skilled performance was characterized through a review of the research on ways experts differ from novices. Next, a series of experiments was conducted to identify differences in knowledge and skill for more and less competent avionics technicians. A difference identified in this experimental work was that skilled technicians exhibited greater proficiency in their understanding of digital logic gates than did less skilled technicians. Skill in reading logic gates appeared to enable effective troubleshooting and, thus, was an important area of study.

Subsequently, Gitomer and Van Slyke (1988) examined technicians' understanding of logic gates through a manual error analysis of the responses of avionics technicians on a logic gate test. Technicians were asked to indicate the output value for 288 logic gates that varied in the type of gates (8), the number of inputs (1,2, or 3), and whether or not inputs were negated. The error analysis identified three classes of errors: 1) technicians who made rule-based errors consistently answered incorrectly problems sharing a set of attributes indicating a misconception in the knowledge needed to solve particular kinds of logic gates. 2) technicians who made weakness area errors had difficulty answering, but did not consistently

answer incorrectly, problems sharing a set of attributes indicating at least an impasse, if not a misconception, in the knowledge needed to solve particular kinds of logic gates, and 3), technicians who made practice area errors made infrequent errors across types of logic gates indicating efficiency could be improved. Some technicians showed more than one class of errors across logic gates with different sets of shared attributes. The error analysis classified 84 of the 119 avionics technicians in the study, or approximately 71 percent, as making rule based and/or weakness area errors Furthermore, over one-third of the sample showed practice area errors. An adaptive instructional system, called GATES, was developed using findings from the error analysis.

Design selection.

The second step in developing PRA is the construction of the observation and measurement designs As I have argued, the validity of both the observation design and the measurement design is evaluated with respect to the substantive base As the work of Gitomer (Gitomer, 1987; Gitomer & Van Slyke, 1988) and his colleagues illustrates, the task for the test developer is to construct and organize observations and combine responses in ways that are consistent with the substantive base. The observation and measurement designs used by Gitomer and Van Slyke (1988) in the GATES tutor are summarized in Figure 3. The tutor performed an initial global assessment followed, if warranted, by a more detailed diagnosis. The global assessment consisted of a circuit tracing task that required tutor users to trace through a complex arrangement of logic gates and indicate the output for each gate[4]. The observation design demanded that logic gates vary in the type of gate, if the gate was negated or not negated, and the number of inputs. Furthermore, the observation design demanded that gates be arranged in a complex circuit. The measurement design demanded that overall accuracy be computed on the task. Tutor users who had high overall accuracy on the circuit tracing task exited the tutor whereas tutor users who answered incorrectly many logic gates attempted a screening test to diagnose the source of their difficulty. Both the observation and measurement design were motivated by substantive concerns. The task involved circuit tracing and was scored using accuracy because substantive research shows that experts and novices in a number of fields, including avionics, differ in their efficiency in accessing domain

knowledge and thus less skilled avionics technicians would have difficulty tracking outputs from logic gates in a complex circuit. Furthermore, technicians who experienced difficulty in tracing through logic gates were administered a further screening test because Gitomer's earlier research indicated that technicians' difficulties interpreting logic gates may be due to misunderstandings, or at least impasses, in those technicians' knowledge.

The global assessment was followed by a screening test in which tutor users indicated the correct output for 48 single gates. The faceted observation design required that gates vary in the type of gate. rather the gate was negated or not negated, and number of inputs. Furthermore, the observation design required that gates be presented singly. The measurement design demanded that accuracy be computed for each gate type, for negated and nonnegated gates, and for gates differing in the number of inputs. Under the measurement design, low accuracy on any set of gates moves the tutor user to a diagnostic module for that set of gates. High accuracy across sets of gates moves the tutor user to a practice module intended to increase the efficiency of the tutor user's access to knowledge of logic gates. Again, both the observation and measurement designs were motivated by substantive concerns. Substantive research suggested two hypotheses for low accuracy on the first module; users may have conceptual difficulty with sets of logic gates or users may have difficulty accessing efficiently logic gate knowledge. Gates that varied in attributes were presented and accuracy on sets of gates was scored to identify conceptual impasse or misunderstanding. Gates were presented singly to reduce the role of efficient access to knowledge in users' performance.

Tutor users who failed to indicate accurately the correct output for one or more sets of logic gates in the screening test were presented with diagnostic modules for those gates. Each module requires tutor users to indicate the correct output for single gates sharing a particular set of attributes. For example, tutor users may be presented with all gates of one type or all negated gates. The observation design required that all gates be presented singly and that gates within a diagnostic module share the same attributes--all of one type, all negated or nonnegated, or all with the same number of inputs. The measurement design required that latent class analysis be used to assign technicians to qualitatively different classes

corresponding to misconceptions regarding logic gates. Using latent class analysis, technicians are assigned to qualitatively different classes by matching response vectors with misconceptions. A latent class approach was used because actual response vectors rarely match exactly an ideal response vector. Matches between actual responses and predicted responses increase support for a particular classification whereas mismatches reduce support for a particular classification.

As with earlier assessment components of the GATES tutor, the observation and measurement designs of the diagnostic modules were motivated by substantive concerns The error analysis of Gitomer and Van Slyke (1988) indicated that three-quarters of technicians may make rule-based errors. The two-stage assessment in which tutor users first completed the screening test and, second, completed, if indicated, diagnostic modules was designed to identify misconceptions in the knowledge needed to solve particular kinds of logic gates. Thus, each diagnostic module consisted of logic gates that shared a set of attributes because the error analysis indicated technicians often held misconceptions in the knowledge needed to solve particular kinds of logic gates. A latent class analysis was used because the error analysis indicated that technicians' misconceptions resulted in systematic patterns of response. Thus, the construction of the assessment components within the GATES tutor is consistent with Gitomer's substantive research on technicians' logic gate understanding.

Test administration.

In the third step of this methodology, test administration, the test developer must consider aspects of administering the test that may influence test takers' performance. The substantive base can inform decisions regarding aspects of test administration such as item or task format, nature of the response, and the context of assessment. The work of Gitomer and his colleagues provides no clear example of substantive considerations in test administration. However, examples from the subtraction domain illustrate aspects of test administration that influence test takers' performance.

For example, substantive research would have much to say about the format, especially the wording, of subtraction word problems intended for children. Generally, a child's poor performance on such problems may indicate a lack of understanding of part-whole relations. However, some researchers argue

that, by the age of 4 or 5, children possess at least tacit understanding of part-whole relations and that what they learn from instruction or through familiarization with problem solving language is how certain verbal formats map onto those relations (De Corte, Verschaffel, & De Win, 1985; Cummins, 1991). Specifically, children at about the first grade level interpret comparative terms as simple possession statements. For example, "Mary has 5 more marbles than John" is interpreted as "Mary has 5 marbles." In addition, the word "altogether" is interpreted as "each". For example, "Mary and John have 5 altogether" is interpreted as "Mary has 5 marbles and John has 5 marbles." As these examples illustrate, test developers would do well to consult substantive research when considering the format of subtraction word problems.

Response scoring.

As the description in Table 4 indicates, response scoring is the implementation of the test theory Practical questions regarding how to manage scoring must be answered and these may be challenging. For example, software must be developed to score and compute individual scores and item and test statistics. I will leave such questions for another occasion.

Instead, I would like to discuss indicators that may be used to evaluate the implementation of the test theory. Generally, CTT implementation is evaluated using item statistics indicating difficulty and discrimination (Millman & Greene, 1989) and test statistics indicating reliability (Feldt & Brennan, 1989). For the purposes of informing instructional decisions, traditional indicators of item and test functioning are not useful. CTT-based statistics evaluate items and tests using the psychological assumptions implicit in test development and the assumption that the test is intended to discriminate which students would profit best from a uniform instructional environment. For example, indicators of reliability assume all items measure the same trait in all test takers. Standards of item functioning emphasize discriminating individuals along a latent continuum. In contrast, PRA-based measures are intended to identify qualitative differences in individuals' skills and knowledge rather than classify individuals along a latent continuum. An immediate need of PRA and other approaches that classify individuals with regards to the knowledge and skill used in solving items or tasks is to develop alternative measures of item characteristics.

Again, the work of Gitomer provides an example. Gitomer and Yamamoto (1989) report the item p values and biserials for 119 technicians who completed a 20 item diagnostic logic gate assessment. As Gitomer and Yamamoto note, difficulty values were moderate and biserials suggested that doing well on one item bodes well for overall test performance. But these item statistics were developed for assessments intended to select students most likely to succeed in a uniform instructional environment rather than assessments intended to make inferences about cognitive structures and processes. What meaning have these statistics for a diagnostic test?

PRA-based assessments require indicators that may be used to evaluate the implementation of test theory that is intended to identify qualitative differences in individuals' skills and knowledge. Researchers have offered several alternatives to traditional CTT indicators of test and item functioning. At the test level, Brown and Burton (1978) propose test diagnosticity as an index to evaluate their diagnostic assessment of students' subtraction misconceptions. Under Brown and Burton's approach, any test taker holding a particular misconception or combination of misconceptions will produce a particular response vector for a given set of test items. The response vectors may be partitioned so that identical response vectors, corresponding to different misconceptions, are placed in the same partition. A perfect diagnostic test would have one response vector in each partition. A less-than-perfect diagnostic test would have at least one partition with more than one response vector.

At the item level, Bart (Bart. 1991; Bart & Williams-Morris, 1990) has proposed two indices. Both indices are computed using an item response-by-rule matrix. For any item, each possible response may correspond to the test takers use of at least one rule or strategy. An item on which the test taker is asked to respond true or false will have two possible responses and each response may correspond to the use of one or more rules. Alternatively, a multiple choice item with four alternatives, such as the two items represented in Table 5, will have four possible responses and each response may correspond to one or more rules. I have represented only four rules in Table 5. Note that an item response corresponding to a rule is represented by a 1 in that cell of the table.

Insert Table 5 about here

The index of response interpretability captures the degree to which each response to the item is interpretable by at least one rule. The computation of response interpretability is straightforward given an item response-by-rule matrix. The index is the number of response that are interpreted by one rule or more divided by the number of responses. Values range from 0 which indicates no rule-based responses, to 1, which indicates complete rule-based responses. In Table 5, the response interpretability of Item 1 is 1 whereas the response interpretability of Item 2 is 0.5.

The index of response discrmination captures the degree to which each response to an item is interptreted by only one rule. Again, the computation of response discrimination is readily understood given an item response-by-rule matrix. Response discrimination is computed in two steps. In the first step 1 is divided by the number of rules that may be used to interpret a response. For example, the first response to Item 1 is interpreted by only one rule In the second step, the sum of the values from the previous step are divided by the number of responses to the item. Values range from 0, which indicates no rule-based responses, to 1, which indicates each response is interpreted by only one rule. In Table 5, the response discrimination for Item 1 is 1 whereas the response discrimination for Item 2 is 0.25.

The PRA-based measures of item and test functioning differ from CTT-based measures in at least two ways: (1) None of the PRA-based indices incorporate the notion of consistently linearly ranking individuals whereas CTT-based discrimination and reliability indices appear to be founded on this notion; (2) All of the PRA-based indices may be computed *a priori* whereas all the CTT-based indices require that the items be administered. These differences reflect differences in emphasis between PRA and CTT test development. PRA development focuses on diagnosing qualitative differences in test takers' knowledge and skill whereas CTT development focuses on selecting individuals most likely to succede in a homogenous educational environment. Furthermore, PRA development emphasizes the test developers' conception of the construct whereas CTT development emphasizes the practical success of the items.

Design review.

Design review is the process of gathering support for the observation and measurement designs used in test development. The design is like a theory, and, as with any scientific theory, the theory is never

proven; rather, evidence is gradually accumulated that supports or challenges the design. Design review is a process that continues before and after test administration. Initially, evidence supporting the design comes from the strength of the research base. Such evidence is similar to Ebel's notion on intrinsic rational validity (Ebel, 1964).

After the test has been constructed and administered, other sources of evidence may be gathered. One source of evidence regarding the validity of the design is the fit between the predictions of the observation design and test takers' performance. The persuasiveness of this sort of evidence lies in the test developer's success modeling how test takers may have solved items or tasks. As Messick (1989) explains:

> Almost any kind of information about a test can contribute to its construct validity . . . Possibly most illuminating of all are direct probes and modeling of the processes underlying test responses, an approach becoming more accessible and more powerful with continuing development in cognitive psychology (pp. 17).

Design revision may be suggested by comparing the results of test administration to the predicted results based on the substantive research used in test development. Anomalous results suggest revisions of the test design and areas of further research. Furthermore, design revision may be suggested by additional substantive research outside of the assessment context. In this way, test development becomes part of basic aptitude and achievement research and "tests can become vehicles of communication between laboratory and field" (Snow & Peterson, 1989, pp. 155).

For example, Gitomer revised the GATES tutor because the latent class measurement design does not associate a value for technicians who made weakness area errors--technicians who had difficulty answering, but did not consistently answer incorrectly, problems sharing a set of attributes. In response to this failure, Gitomer and Yamamoto (1991) applied Yamamoto's (1989) HYBRID model to assess technicians' understanding of logic gates. Using the HYBRID model, "Individuals whose responses are not consistent with one of the LCM [latent class model] classes may be modeled more conservatively by a continuous model that makes no strong assumptions about qualitative understanding but simply quantifies

their overall level of proficiency" (pp175).

Conclusions on a New Assessment Approach

I was trying to understand better differences between someone who used analogy well and someone who used analogy poorly when I examined first the responses of students to figural analogies. I was a psychologist attempting to model an important feature of learning--the use of analogy. I realized later that I was also gathering validity evidence for an analogical reasoning test. I soon concluded that test developers could design assessments through systematic, research-based variations in problem characteristics. Conversely, test developers could infer the processes and knowledge structures used by an individual through systematic, research-based patterns of problem responses. Apparently, many other researchers have come to that same conclusion. Gradually, the distinction between psychology and psychometrics has become obscured. But many issues must be addressed before PRA can be considered as viable an assessment methodology as the time-tested CTT or even the relatively new IRT. I address several of those issues in this final section.

What is the role of substantive findings in applications?

Substantive theory and research play a prominent role in PRA. The substantive base is the foundation of PRA. Given the prominence of substantive theory and research in PRA, the assumptions concerning its role should be examined closely. An assumption underlying PRA is that substantive theory and research may be applied usefully to construct assessments. Critics may object that theory and research from the laboratory have little to say about tasks and variables in real world domains that are the focus of instruction. This argument appears to misrepresent cognitive science as an enterprise situated solely in the laboratory. As the work in everyday memory (Loftus, 1991) and situated cognition (Lave, 1988) illustrates, cognitive scientists devote much time to understanding how people represent and process information in real world domains. The work of VanLehn, Tatsuoka and Gitomer reviewed in this paper deals with individual differences in real world domains.

Another assumption underlying PRA is that assessment results may be used to test theory and to inform research. Critics may raise several objections against this assumption. Critics may object that

experiment is the only valid way to test theory. True, experiment provides powerful tests of theory but not the only tests (Bickhard, 1992). Experimental tests in many fields, including geology, evolutionary biology, and astronomy, are generally not possible.

Critics may object that alternative theories may account for the assessment results and so successful PRA applications have no implications for the specific substantive base used to generate the assessment. In the social sciences, I doubt there is any single data set that cannot be explained in multiple ways in a post hoc fashion. Note that the substantive base in PRA is used to generate assessments and predict results. Any number of explanations are plausible when theorizing is post hoc but fewer theories are successful in predicting results.

Who makes these new assessments?

Under the PRA approach, what kind of expertise must the test developer possess? The emphasis in PRA given to multiple disciplines appears to demand an encyclopediast. However, few individuals are able to master the knowledge and skills required to be expert in many areas. A solution may be the use of a team to develop PRA (Tittle, 1991). The team would include cognitive scientists to address psychological issues, statisticians to address statistical issues, and subject matter experts to insure the accuracy of substantive issues. The role of the team leader would be to coordinate these different aspects of test development and so the team leader must be familiar with the issues in each area. The position of team leader may be filled best by an educational psychologist or other professional whose training includes educational, psychological and statistical concerns.

In this paper, I have outlined several points on which PRA differs from traditional measurement approaches. However, PRA and traditional measurement are similar in the sense that a quality assessment depends on a careful, creative, and conscientious test development staff. The development of a PRA is not straightforward and such a conclusion should not be drawn from the step-by-step description of test development. Even an assessment generated using an algorithm has an element of art. The success of a PRA depends in large part on the competence of the test development staff.

Does assessment drive instruction or instruction drive assessment?

I will conclude this paper with a short, and incomplete, discussion of the relationship between diagnostic assessment and instruction. As I noted earlier, work on PRA has been motivated, at least in part, by the current emphasis on helping individuals to succeed in educational opportunities in contrast to selecting individuals for those opportunities. Therefore, an important aspect, perhaps the most important aspect, in developing diagnostic assessments is to link assessment results to instructional alternatives The lack of such a link currently has produced assessment detached from instruction. For example, what does it mean for instruction to have children score below the 50th percentile on reading achievement? However, the detachment of assessment from instruction will always exist unless assessment and instruction are based on common understandings of learning and achievement.

Some proponents of diagnostic assessment appear to assume that these new assessments will necessarily improve instruction and learning. However, the impact of assessment results depends on the function the results are intended to play. Assessment plays different functions in education (Resnick & Resnick, 1992): (1) public accountability and program evaluation; (2) instructional management and monitoring; (3) student selection and certification. Each function is intended for a different audience and requires assessments with different characteristics. Much of the PRA research is attempting to develop assessment for instructional management and monitoring intended to guide teachers and students in daily classroom work. The educational community must determine if assessment for instructional management and monitoring is a worthwhile use of students' time or may be done more effectively by the teacher in the traditional classroom setting. Diagnostic assessment can play other functions in education including certifying competence and evaluating programs--assessments that reveal the cognitive mechanisms test takers use provide important information on whether instruction is achieving its goal. A better understanding of the ways diagnostic assessment can improve instruction and learning would be useful for researchers and policy makers who must decide what issues to address in diagnostic assessment.

References

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

Anastasi, A. (1967). Psychology, psychologists, and psychological testing. American Psychologist, 22, 297-306.

Anastasi, A. (1982). Psychological testing (5th ed.). New York: Macmillan.

Bart, W. M., & Williams-Morris, R. (1990). A refined item digraph analysis of a proportional reasoning test. Applied Measurement in Education, 3, 143-165.

Bart, W. M. (1991, April). A refined item digraph analysis of Siegler's balance beam tasks. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Beatty, L. S., Madden, R., Gardner, E. F., & Karlsen, B. (1976). Stanford Diagnostic Mathematics Test: Manual for administering and interpreting. New York: Harcourt Brace Jovanovich.

Bejar, I. I. (1984). Educational diagnostic assessment. Journal of Educational Measurement, 21, 175-189.

Bickhard, M. H. (1992). Myths of science: Misconceptions of science in contemporary psychology. Theory & Psychology, 2, 321-337.

Brennan, R. L. (1992). Elements of generalizability theory(2nd ed.). Iowa City, IA: American College Testing.

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

Carnegie Council on Adolescent Development (1989). Turning points: Preparing American youth for the 21st century. Washington, D.C.: The Carnegie Corporation of New York.

Carroll, J. B. (1974). The aptitude-achievement distinction. The case of foreign language aptitude and proficiency. In D. R. Green (Ed.), The aptitude-achievement distinction. Monterey, CA: CTB/McGraw-Hill.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Cummins, D. D. (1991). Children's interpretations of arithmetic word problems. Cognition and Instruction, 8, 261-289.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. Psychological Bulletin, 81, 95-106

De Corte, E., Verschaffel, L., & De Win, L. (1985). The influence of rewording verbal problems on children's problem representation and solutions. Journal of Educational Psychology, 77, 460-470.

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

Feldt. L. S., Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 105-146). New York: Macmillan.

Gitomer, D. H. (1984). A cognitive analysis of a complex troubleshooting task. Unpublished doctoral dissertation, Pittsburgh, PA: University of Pittsburgh.

Gitomer, D. H. (1987, October). Using error analysis to develop diagnostic instruction. Paper presented at the meeting of the Military Testing Association.

Gitomer, D. H., & Van Slyke, D. A. (1988). Error analysis and tutor design. International Journal of Machine Mediated Learning, 2, 333-350.

Gitomer, D. H., & Yamamoto, K. (1989, April). Using embedded cognitive task analysis in assessment. Paper presented at the annual meeting of the American Educational Research Association.

Gitomer, D. H., & Yamamoto, K. (1991). Performance modeling that integrates latent trait and class theory. Journal of Educational Measurement, 28, 173-189.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. American Psychologist, 36, 923-936.

Glass, G. V. (1986). Testing old, testing new: Schoolboy psychology and the allocation of intellectual resources. In B. S. Plake & J. C. Witt (Eds.), The future of testing (pp. 9-28). Hillsdale, NJ: Erlbaum.

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

Ippel, M. J. (1986). Component-testing: A theory of cognitive aptitude measurement. Amsterdam, The Neatherlands: Free University Press.

Ippel, M. J. (1991). An information-processing approach to item equivalence. In P. L. Dann, S. H. Irvine, & J. H. Collis (Eds.), Advances in computer-based human assessment (377-396). Boston: Kluwer.

Langley, P., Wogulis, J., & Ohlsson, S. (1990). Rules and principles in cognitive diagnosis. In N. Frederiksen, R. Glaser, A. Lesgod, & M. G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 217-250). Hillsdale, NJ: Erlbaum.

Lave, J. (1988). Cognition in practice. Boston: Cambridge.

Linn, R. L. (1986). Educational testing and assessment: Research needs and policy issues. American Psychologist, 41, 1153-1160.

Loftus, E. F. (1991). The glitter of everyday memory...and the gold. American Psychologist, 46, 16-18.

Lohman, D. F. (in press). Component scores as residual variation (or why the intercept correlates best). Intelligence.

Lohman, D. F., & Ippel, M. J. (in press). Cognitive diagnosis: From statistically-based assessment toward theory-based assessment. In N. Fredriksen, R. J. Mislevy, & I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum.

Lohman, D. F., & Nichols, P. D. (1990). Training spatial abilities: effects of practice on rotation and synthesis tasks. Learning and Individual Differences, 2, 67-93.

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

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental tests scores. Reading, MA: Addison-Wesley.

McNemar, Q. (1964). Lost: Our intelligence? Why? American Psychologist, 19, 871-882.

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement, 21, 215-237.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-104). New York: American Council on Education and Macmillan.

Mislevy (in press). Foundations of a new test theory. In N. Fredriksen, R. J. Mislevy, & I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum.

National Education Goals Panel (1991). Measuring progress towards the national education goals: Potential indicators and strategies. Washington. DC: National Education Goals Panel.

National Governors' Association (1990). Educating America: State strategies for achieving the national education goals. Washington, D.C.: National Governors' Association.

Nichols, P. (1990). Cognitive assessment of figural analogical reasoning: A theory-driven approach toward test development. Unpublished doctoral dissertation, Iowa City, IA: University of Iowa.

Nichols, P. (1991, April). The psychology of item construction. In D. F. Lohman (Chair), Constructing test items and tasks using an understanding of the psychology of the domain. Symposium conducted at the annual meeting of the American Educational Research Association, Chicago, IL.

Ohlsson, S. (1990). Trace analysis and spatial reasoning: An example of intensive cognitive diagnosis and its implications for testing. In N. Frederiksen, R. Glaser, A. Lesgod, & M. G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 251-296). Hillsdale, NJ: Erlbaum.

Pellegrino, J. W. (1992). Commentary: Understanding what we measure and measuring what we understand. In B. R. Gifford & M. C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement, and instruction (pp.275-300). Boston, MA: Kluwer Academic Publishers.

Pohl, H. L., & Nutter, J. T. (1985). Use of analogy in computer language acquisition. AEDS Journal, 18, 254-266.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), Changing assessments; Alternative views of aptitude, achievement and instruction (pp. 37-75). Boston, MA: Kluwer.

Self, J. A. (1990). Bypassing the intractable problem of student modeling. In C. Frasson & G. Gauthier (Eds.), Intelligent tutoring systems; At the crossroad of artificial intelligence and education (pp. 107-123). Norwood, NJ: Ablex.

Shepard, L. A. (1991). Psychometricians' beliefs about learning. Educational Researcher, 20, 2-9.

Sleeman, D., Kelly, A. E., Mortinak, R., Ward, R. D., & Moore, J. L. (1989). Studies of diagnosis and remediation with high school algebra students. Cognitive Science, 13, 551-568.

Sleeman, D. H., Langley, P., & Mitchell, T. M. (1982, Spring). Learning from solution paths: An approach to the credit assignment problem. AI Magazine, pp. 48-52.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 263-332). New York: Macmillan.

Snow, R. E., & Mandinach, E. B. (1991). Integrating assessment and instruction: A research and development agenda (RR-91-8). Princeton, NJ: Educational Testing Service.

Snow, R. E., & Peterson, P. (1985). Cognitive analyses of tests: Implications for redesign. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics (pp. 149-166). New York: Academic Press.

Snow, R. E., & Yalow, E. (1982). Education and intelligence. In R. J. Sternberg (Ed.), Handbook of human intelligence (pp. 493-559). Cambridge, MA: Cambridge University Press.

Spiro, R. J., Feltovich, P. J., Coulson, R. L., & Anderson, D. K. (1989). Multiple analogies for complex concepts: Antidotes for analogy-induced misconceptions in advanced knowledge acquisition. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 498-531). Cambridge, MA: Cambridge University Press.

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

Stiggins, R. J. (1991). Facing the challenges of a new era of educational assessment. Applied Measurement in Education, 4, 263-273.

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

Thorndike, R. L. (1975). Mr. Binet's test 70 years later. Educational Researcher, 4, 3-7.

Tittle, C. K. (1991). Changing models of student and teacher assessment. Educational Psychologist, 26, 157-165.

VanLehn, K. (1982). Bugs are not enough: Empirical studies of bugs, impasses, and repairs in procedural skills. Journal of Mathematical Behavior, 3, 3-72.

VanLehn, K. (1988). Student modeling. In M. C. Polson, J. J. Richardson, E. Soloway (Eds.), Foundations of Intelligent Tutoring Systems (pp. 55-78). Hillsdale, N.J.: Erlbaum.

Wenger, E. (1987). Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge. Los Altos, CA: Morgan Kaufman.

Yamamoto, K. (1989). Hybrid model of IRT and latent class models (ETS Research Report No. RR-89-41). Princeton, NJ: Educational Testing Service.

Footnotes

[1]Other terms were considered and rejected. Theory-referenced construction has been used but was rejected because psychometric theories have been used in the past and this could be called theory referenced construction.

[2]IRT ability estimates and CTT number correct scores typically correlate .95 (Mislevy, in press).

[3]Technicians must be able to read and interpret logic gate symbols, common components of schematic diagrams, in troubleshooting electronics equipment.

[4]The global assessment included a circuit troubleshooting task that was not scored and so is not discussed.

Table 1

Items from a diagnostic test of multicolumn subtraction (from VanLehn, 1982).

	647	885	83	8305	50	562
	- 45	-205	-44	- 3	-23	- 3
RULE A	602	680	41	8302	33	561
RULE B	602	680	39	8302	27	559
	742	106	716	1564	6591	311
	-136	- 70	-598	- 887	2697	-214
RULE A	614	176	282	1323	4106	163
RULE B	606	36	118	677	3894	97
	1813	102	9007	4015	702	2006
	- 215	- 39	-6880	- 607	-108	- 42
RULE A	1602	137	3887	4612	606	2004
RULE B	1598	63	2227	3408	604	2064
	10012	8001				
	- 214	- 43				
RULE A	10202	8042				
RULE B	10898	8068				

Table 2

Operations for subtraction (from Langley, Wogulis, & Ohlsson, 1990).

Add-Ten (number, row, column) Takes the number in a row and column and replaces it with that number plus ten.

Decrement (number, row, column) Takes the number in a row and column and replaces it with that number minus one.

Find-Difference (number1, number2, column) Takes the two numbers in the same column and writes the difference of the two as the result for that column.

Find-Top (column) Takes a number from the top row of column and writes that number as the result for that column.

Shift-Column (column) Takes the column which is both focused-on and being processed and shifts both to the column on its left.

Shift-Left (column) Takes the column which is focused-on and shifts the focus of attention to the column on its left.

Shift-Right (column) Takes the column which is focused-on and shifts the focus of attention to the column on its right.

Table 3

Item objectives for the Subtraction of Whole Numbers Concept/Skill Domain in the SDMT.

The pupil will demonstrate the ability to use the standard algorithm for subtraction (vertical form) with renaming by:

Item	
4	finding the unknown addend (remainder) when a number in the tens is subtracted from a number in the hundreds.
5	finding the unknown addend when a number in the hundreds is subtracted from another number in the hundreds.
6	finding the unknown addend when a number in the hundreds is subtracted from a number in the thousands.
7	finding the unknown addend when a number in the hundreds is subtracted from a number in the thousands with a zero in the tens place.
8	finding the unknown addend when a number in the thousands is subtracted from another number in the thousands with a zero in the ones and in the hundreds places.
9	finding the unknown addend when a number in the thousands is subtracted from another number in the thousands with a zero in the tens place.

Table 4

The seven steps of test development within psychology-driven test development.

STEP 1.	SUBSTANTIVE THEORY CONSTRUCTION The substantive base concerns the development of a model or theory that describes the knowledge and skills hypothesized to be involved in performance and the item or task characteristics hypothesized to interact with the knowledge skill.
STEP 2.	DESIGN SELECTION In this step, the test developer selects the observation and measurement designs. The selection is informed by the substantive base constructed in step 1. Subsequently, the test developer constructs items or tasks that will be responded to in predictable ways by test takers with specific knowledge, skills, and other characteristics identified as important in the theory. The procedure for constructing assessments is the operationalisation of the assessment design.
STEP 3.	TEST ADMINISTRATION Test administration includes every aspect of the context in which test takers complete the test: the format of the items, the nature of the required response, the technology used to present test materials and record responses, and the environment of the testing session. Decisions concerning the context of the testing session should informed by research on how aspects of the context influences test takers' performance.
STEP 4.	RESPONSE SCORING The goal of this step is to assign values to test takers' patterns of responses so as to link those patterns to theoretical constructs such as strategies or malrules. As with assessment construction, a scoring procedure is the operationalize of the assessment design.
STEP 5.	DESIGN REVISION Design revision is the process of gathering support for a model or theory. As with any scientific theory, the theory used in test development is never proven; rather, evidence is gradually accumulated that supports or challenges the theory. In this step, the results of administering the assessment are used to revise the substantive base upon which was based the construction of the assessment.

Table 5

The Item Response by Rule matrix for two multiple choice items.

		RULE			
	RESPONSE	1	2	3	4
ITEM 1	1	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1
ITEM 2	1	0	0	1	1
	2	0	0	0	0
	3	0	0	0	0
	4	1	1	0	0

Figure Captions

Figure 1. Examples of fraction subtraction problems (from Tatsuoka, 1990).

Figure 2. Two elements of a substantive theory.

Figure 3. A representation of the observation and measurement design of the GATES tutor.

EXAMPLE OF "ADD 10" PROCEDURE

$$2\frac{5}{10} - \frac{6}{10} = 1\frac{15}{10} - \frac{6}{10} = 1\frac{9}{10}$$

ITEMS THAT DISCRIMINATE THE MISCONCEPTION

1. $2\frac{5}{10} - \frac{6}{10} = ?$

2. $2\frac{2}{5} - \frac{1}{5} = ?$

Figure 1. Examples of fraction subtraction problems (from Tatsuoka, 1990).

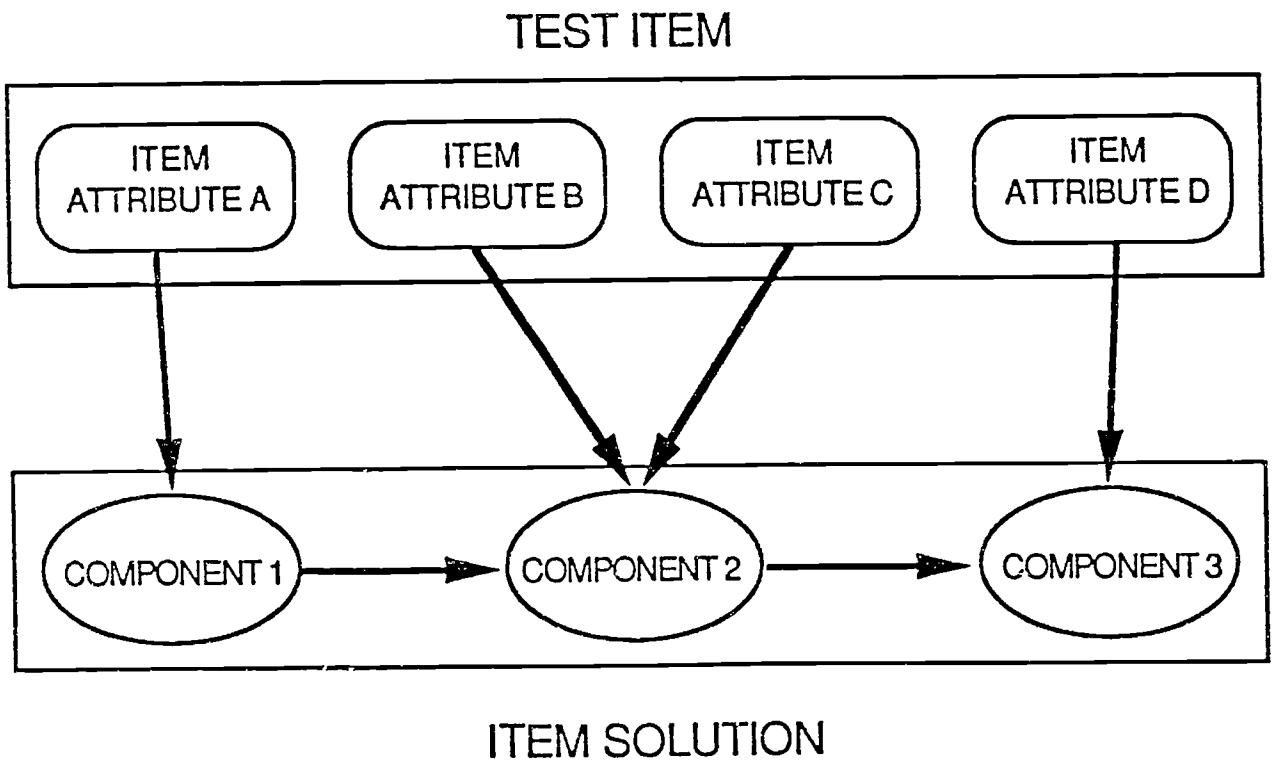

Figure 2. Two elements of a substantive theory.

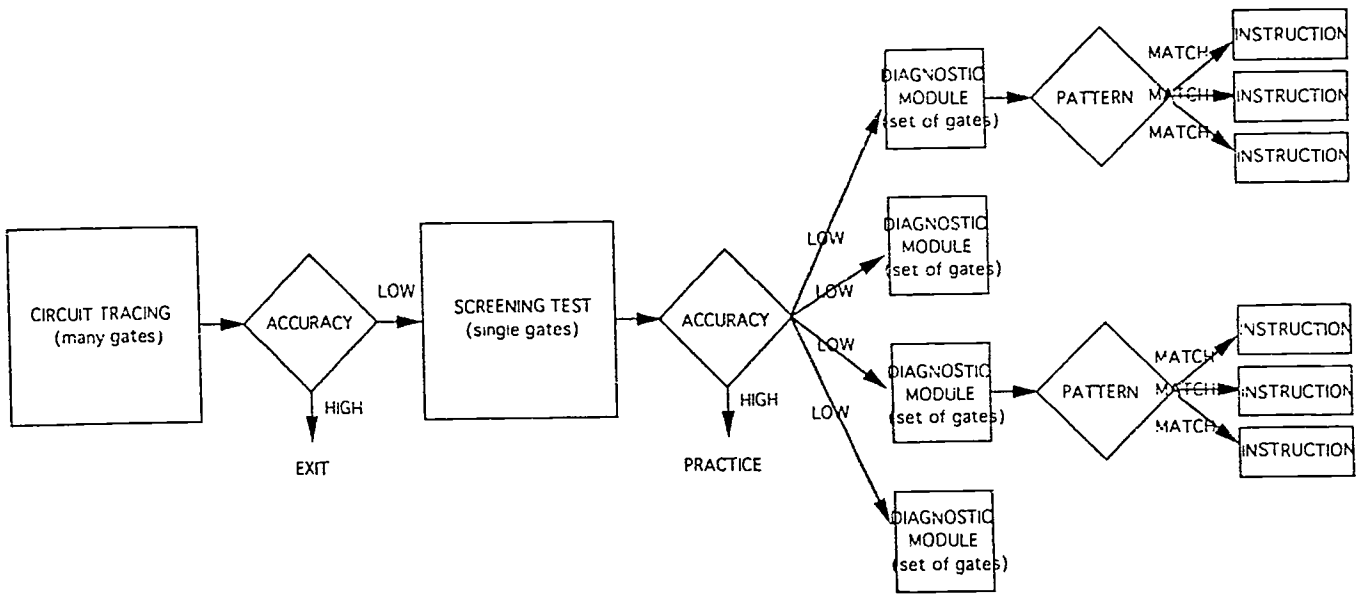

Figure 3. A representation of the observation and measurement design of the GATES tutor.