# Psychometricians' Beliefs About Learning

LORRIE A. SHEPARD

*The author contends that disputes within the measurement community about what constitutes legitimate test preparation and whether "teaching to the test" is good or bad for student learning can be explained by differences in measurement specialists' beliefs about learning. Qualitative analysis of interview data from a nationally representative sample of 50 district testing directors revealed that approximately half of the measurement specialists operate from implicit learning theories that advocate, first, close alignment of tests with curriculum and, second, judicious teaching of tested content. Historical quotations are used to show that these beliefs, associated with criterion-referenced testing, derive from behaviorist learning theory, which requires sequential mastery of constituent skills and explicit testing of each learning step. The sequential, facts-before-thinking model of learning is contradicted, however, by a substantial body of evidence from cognitive psychology. Implicit beliefs should be made explicit because an understanding of learning theory assumptions is fundamental to evaluating evidence of testing effects and therfore to framing validity investigations.*

*Educational Researcher, Vol. 20, No. 6, pp. 2–16*

In this article I examine beliefs that psychometricians hold about learning. What conceptions of teaching and learning do measurement specialists invoke when they make decisions about testing practice? In proposing this line of inquiry, I borrow both methodological approach and perspective from recent research on teacher thinking, which suggests that teachers' classroom practices can be understood in terms of their beliefs or implicit theories about instruction and learning. As described by Clark (1988), "These theories are not neat and complete reproductions of the educational psychology found in textbooks or lecture notes. Rather, teachers' implicit theories tend to be eclectic aggregations of cause–effect propositions from many sources, rules of thumb, generalizations drawn from personal experience, beliefs, values, biases, and prejudices" (p. 6). Similarly, psychometricians likely hold both shared and idiosyncratic ideas about student learning and the role of testing in effective instruction.

The possibility that measurement specialists have unstated learning theories that influence their practices of testing and assessment is suggested by several observations. For example, in telephone interview data from state directors of testing, there was almost uniform agreement among the 40 directors who characterized their testing programs as having "high stakes" that high-pressure tests focused more instructional time and attention on tested objectives (Shepard, 1990a). However, respondents differed as to whether they attached a positive or negative "valence" to the teaching changes they perceived in response to testing. By implication some believed that students in their state would learn *more* because high-stakes testing forced attention to important skills that had hitherto been neglected. In contrast, those who worried about the effects of testing on instruction believed that somehow something would be lost if the tests reshaped curriculum. These two groups did not appear to differ by the amount of reported pressure associated with testing nor by the type of test administered (i.e., norm-referenced survey test or a test designed to be objectives referenced; thus it was more plausible to infer that differences in belief systems accounted for differences in respondents' interpretations of effects.

A similar difference in perspective can be seen in arguments about what constitutes legitimate test preparation. Mehrens and Kaminski (1989) conducted a content analysis of one version of the test preparation materials called *Scoring High* and found them to be so similar to the actual test that, in the judgment of Mehrens and Kaminski, using these materials would be the same as practicing with the test beforehand and therefore unethical. Makers of *Scoring High*, however, recommend that their materials be used daily for 4–5 weeks before regularly scheduled standardized testing (*Scoring High on the Iowa Test of Basic Skills*, 1987). They assert that their materials uphold the principles of the Code of Fair Testing Practices in Education, (Joint Committee on Testing Practices, 1988) by identifying learning gaps and removing sources of irrelevant difficulty by familiarizing children with test formats (American School Publishers, 1989). This dispute can be framed in traditional terms of test validity, but it can also be construed as a dispute about how learning occurs. The antagonists likely differ in their beliefs about transfer of training from specific tasks, the role of practice and repetition, and the desirability of using multiple-choice formats for first-time instruction.

Lastly, the debate between Popham (1987) and Bracey (1987) or Popham (1987) and Shepard (1988) about the efficacy of measurement-driven instruction is motivated by conflicting learning theories. It is not just that we disagree about unintended side-effects of measurement-driven instruction, as when tested content grows to command more and more instructional time. Bracey, Shepard, and others disagree fundamentally with measurement-driven basic-skills instruction because it is based on a model of learning

LORRIE A. SHEPARD *is Professor in the School of Education, Campus Box 249, University of Colorado, Boulder, CO 80309. Her specializations are educational measurement and policy research.*

which holds that basic skills should be taught and mastered before going on to higher order problems, as Popham suggests when he says, "Creative teachers can *efficiently* promote mastery of content-to-be-tested and then get on with other classroom pursuits" (p. 682).

It is my contention, prompted by recurring themes in earlier data from state testing directors (Shepard, 1990a), that these differences of opinion about the role of testing and accompanying assumptions about learning are not limited to a few authors writing in the measurement literature, but rather reflect a major divide in the measurement community at large. I contend further that these differences in theoretical perspectives explain why some measurement specialists are pleased by more teaching to tested content while others are not and that these differences in beliefs can account for differences in practices such as the type or extent of test preparation. The present study provides a more systematic examination of measurement specialists' beliefs with an independent sample of specialists. It is organized into four parts: (a) an analysis of interview data from a nationally representative sample of 50 district testing directors; (b) a comparison of test directors' conceptions of learning with the frameworks of criterion-referenced testing, programmed instruction, and behaviorist psychology; (c) consideration of a competing learning model from cognitive psychology; (d) implications of explicit understandings of learning theory for reform of assessment practice.

## Implicit Learning Theories: Interviews With 50 District Test Specialists

### *Data Source*

The interview transcripts examined here were collected as part of a larger study to replicate and extend Cannell's (1987) controversial report which asserted that all 50 states and 90% of U.S. school districts claim to be above average. Test data from the 35 states with normative statistics and from 153 districts (responding from a stratified random sample of 175) were reported in Linn, Graue, and Sanders (1990). The Linn et al. technical report also describes the method of sampling districts by region, size, and socioeconomic strata and includes the original survey instruments, both mailed questionnaires and telephone interview protocols. As described in Shepard (1990a), telephone interviews were conducted with the directors of testing from all of the 50 states regarding the uses of test data, the process of test selection, time spent on teaching tested objectives, objectives given less time as a result of the test, guidelines for test preparation, typical and extreme practices in preparing students to take tests, and test security efforts and experiences. Parallel telephone interviews, which provided the data examined here, were also conducted with a subsample of 50 district test directors. Methods by which the district subsample was selected to be representative of the national population of school districts are described in Appendix E of Linn et al.

### *Data Analysis*

Although test directors' elaborations about the purpose of testing, and indirectly their assumptions about learning or instruction, sometimes occurred in answer to any of the interview questions, three prompts were selected for systematic reanalysis because these questions most often elicited talk about the effects of testing on instruction and learning. As shown in Appendix A, Questions 15, 16, and 17 asked whether efforts had been made to ensure that the curriculum and district (or state) test were aligned, whether teachers spend more time teaching the specific objectives on the tests than they would if the tests were not required, and whether important objectives are given less time or emphasis because they are not included on the test.

After responses to Questions 15–17 were read separately and counted as yes, no, or don't know, interview transcripts for the question sets were reread and characterized by a phrase or sentence to reflect each respondent's overall opinion about the effect of mandated tests on instruction in the district. Similar responses were then grouped together to form categories. To facilitate the initial sorting task (i.e., to check for similarity within category and meaningful distinctions between categories) and later as a reporting device, categories were arranged along a continuum from least to greatest test influence on instruction. Although the initial reading and summarization of state interviews (for Shepard, 1990a) had suggested two other possible categorization schemes—views about criterion-referenced testing and learning or positive versus negative opinions about testing impact—the decision was made to organize the data in terms of the degree of instructional influence of tests because this scheme stayed closest to the survey as posed and therefore required the least inference on the part of the coder. This continuum also accounted for all of the data, whereas the other schemes omitted some cases which could not be accurately categorized. In keeping with the decision to stay close to the data for initial analysis, responses were located on the continuum according to the explicit answer choices of the respondents. Often a test director would describe a situation which implied substantial influence of tests on instruction to the interviewer or to the reader; nonetheless, efforts were made to categorize responses from the perspective of the respondent. This procedure sometimes led to different categorizations for highly similar accounts. For example, in Appendix A, Test Director 9 in Category II and Test Director 15 in Category IV gave very similar answers about the tendency for teachers to pay attention to tested objectives and about district efforts to make sure that teachers attend to important objectives beyond those tested. They differed, however, in their explicit answers to Question 16, with only one saying that more time was spent teaching tested objectives, and were therefore assigned to different categories.

Quantitative and qualitative data displays were developed. Brief phrases were used to convey different meanings for yes-no responses. Paraphrased quotations were developed to represent the gist of each category. Then shortened quotations were selected to provide specific examples of the types of answers given in each category.[2]

## Inferences About Implicit Learning Theories

Clearly measurement specialists in these two samples were not asked directly about their beliefs or theories of learning. Inference is required to hear assumptions about learning in talk about the effects of testing. Although this mode of investigation is not as concrete as some would like, it is customary to use indirect means to study the implicit theories of practitioners, given that nonexperts are not expected to have their theories easily accessible to report in propositional form. (Although test directors have expertise about measurement, they do not usually consider themselves to be experts about

learning theory.)

Interpretations about what measurement specialists believe about learning are based on reanalysis of the primary narrative data. Again descriptive codes were used to typify the responses. Those codes eventually became the propositional summaries used here to present the data. The data were reread for counterexamples. In general, the data did not produce equally elaborated competing theories of learning. Instead, one persistent model which seems to be widely shared in the profession emerged; I called it *the criterion-referenced-testing learning theory*. A competing perspective, much less well elaborated in terms of an underlying learning model, was also identified which might be called *the anti-measurement-driven instruction* position. As stated previously, some cases could not be categorized accurately at this higher level of inference. Therefore, beliefs about learning are presented below as propositions followed by supporting quotations and estimates of the proportion of cases accounted for. The first two propositions characterize the criterion-referenced-testing learning theory perspective. By way of contrast, the third proposition summarizes the more loosely defined antimeasurement-driven instruction position.

1. *If a test is "criterion-referenced" or "curriculum-referenced," it is desirable for instructional effort to be redirected toward the test.* The term *criterion-referenced* test is in quotation marks because test directors often referred to tests keyed to important instructional objectives as representing the appropriate goals of instruction even when they were off-the-shelf standardized norm-referenced tests. Thus I am using the term to characterize their way of speaking about the use of a test matched to important objectives whether or not they used the term explicitly. Even when a test was officially called a criterion-referenced test, it should be noted that this now appears to mean a test developed by tying items to instructional objectives. "Criterion-referenced" has become synonymous with content-referenced or objectives-referenced. In the vernacular it does not mean external referencing of test scores to criterion performances as suggested by Glaser's (1963) original use of the term.

Two entire categories of responses on the instructional effects dimension in Appendix A represent "criterion-referenced-testing" types, Category III and Category V. Both groups reported a great deal of instructional effort addressed to tested objectives and emphasized that these were the important objectives that should be taught. Respondents in Category III, however, denied that this focusing required any redirecting of attention from what would have been taught if the test were not used.

Criterion-referenced-testing rhetoric is epitomized by Respondent III.7:

> We have a locally developed criterion referenced testing program, and these are skills that we have identified as being absolutely essential, and we test and retest until students show mastery. This is the kind of test that we think teachers should teach to, not particular items and answers of course, but really focus on the curriculum, because we have identified (___) as key.

In other words, the tests and the curriculum are synonymous. Test Director III.8 speaks in the same criterion-referenced terms about the standardized norm-referenced test in use in his district for the past 10 years:

> [16: More time teaching tested objectives?] No. I think that most of the skills that are appraised in the assessment instruments are part of our curriculum. They've always been part of the curriculum. When we're talking about skills, they've been there. I think pretty much the assessment instruments match what skills have been taught and are being taught.

Likewise any of the quotations in Category V can be used as examples of a learning model which says something like the following: "In order for children to learn effectively in schools, the schools must have a well specified set of objectives, accountability tests should be keyed to these essential skills, and feedback should be provided about how well students have mastered the desired objectives." For example, Respondent V.15 stated:

> [16: More time teaching tested objectives?] Probably. They don't teach to items. We don't give them item analysis. We give them an integrated report grouped by domain. For example, for dealing with reading comprehension, we would have broken that down through a computer to facts or opinion, to main idea, to details, to sequence, to generalization. They would not see individual items. So they teach to those areas. Those areas, in turn, are curriculum referenced, and there are support materials for all of them.

Categories III and V account for 28% of all the district test directors. In addition, approximately half of the respondents in Categories IV and VI also gave positive accounts of a test carefully matched to the curriculum which improved instruction by directing attention to important objectives. For example, on the basis of separate analysis of Question 15, twenty-six of the 50 test directors described extensive efforts to bring curriculum and teaching in line with the test. The following quotations are answers to Question 15 selected to represent those who espouse a criterion-referenced view of test–curriculum alignment from among respondents in Categories IV and VI. (Original identification codes are used when the case was not one of the illustrative cases in Appendix A.)

> IV.[1722]: Yes, very extensive. With regard to the state test...there was a major effort to do a curriculum match between the content of the state test and the curriculum of the school district.

> IV.[1841]: If I can use a term that's often used by (___), we are very much involved in a test-driven curriculum, right or wrong. As we look at what the tests are attempting to measure, we have made adjustments in our curriculum to make sure that those pieces are in fact being covered.

> VI.23: Yes, there have been strands and objectives which have been prepared for [city] which would identify those strands and objectives which are measured by the CAT, also by our [state] test. So there would be correlations that have been developed for both of these tests to identify those areas and to provide techniques or lessons or methods that would help teachers obtain these objectives in classes.

To restate, then, test directors who think about learning from a criterion-referenced-testing perspective believe that it is appropriate and desirable for the test to be the target for instruction. This perspective is shared by approximately half of the sample of district test directors, many of whom were describing a local or state use of a norm-referenced test rather than

a test designed specifically as a criterion-referenced measure.

2. *Basic skills are the most important learning goals, especially for elementary education, because basic skills are the building blocks or prerequisites for subsequent learning.* Instances of the "basic skills" proposition were less frequent and tended to be embedded within the protocols already associated with Proposition 1. The following excerpts are illustrative of the perspective that learning objectives should be sequenced to ensure mastery.

> V.14: But if you're attempting to ready kids for the achievement test, you're attempting to ready students for the curriculum tests that are developed within the local efforts. Then that could take most of the time.... But when you say less important [Question 17], I don't know. The things that we try to stress are what is important. And of course you have terminal objectives and supporting objectives. But to push the terminal objectives which one might consider important, you have to in many respects touch upon the building-block objectives.

> V.19: Well, it is a criterion referenced test, the [state test] that I mentioned, and all of those skills are remediated, taught and then remediated after the test at every grade level, and that is its purpose, because by the time they get to be in high school prior to graduation, they must have mastered them. In order that the courts would allow us to withhold a diploma, we had to give evidence that we are teaching those skills adequately.

> V.20: We have what we call the basic elements of our curriculum, and our [local tests] reflect those basic elements. [State test aligned?] As closely as we can get it. That sometimes is a problem, but by and large, the state has made quite an effort in the last 4 or 5 years to get everybody in line for at least minimum skills or basic skills.... I don't believe the test eliminates any really important objectives.

Occasionally respondents who had not previously been classified as having a criterion-referenced testing perspective referred to the importance of teaching essential skills. For example, one declared:

> IV.12: So they established this list of essential skills. It took about a year to do that for each grade and each of those subject areas, what ought to be taught, the essential skills that ought to be taught at each grade level. And once we received these, we made sure that every teacher and administrator in our district had a copy of these, and they were instructed to make sure that they taught all of these essential skills at their particular grade level.

Together Propositions 1 and 2 comprise what I have called the criterion-referenced-testing learning theory. These themes or shared understandings which seemed to recur in the first reading of the data were the impetus for this analysis. More systematic investigation confirms that many measurement specialists have a coherent view of learning as the sequential mastery of basic skills. Testing is closely tied to instruction because it assesses what students know and don't know in their progress toward mastery. This underlying learning model is elaborated further in the next section of the article by examining the work of psychologists from whom measurement specialists appear to have drawn their assumptions about learning.

To complete this second-level analysis, however, where learning theories are inferred from narratives about instructional effects of testing, I offer one final belief or proposition which accounts for most of the cases not characterized by the criterion-referenced testing perspective.

3. *Tests should be for monitoring but should not drive instruction.* As stated previously, whatever learning beliefs are held by those who do not believe in the criterion-referenced-testing learning theory, they were not adequately elicited by these indirect questions on the instructional effects of testing. That is, in the course of telling whether they believed that tests in their jurisdiction had or had not increased the amount of time spent teaching specific objectives, they did not reveal as much about their *learning theory* as the criterion-referenced group had. Perhaps this asymmetry in the adequacy of the data occurred because large-scale testing and learning are closely tied together only from the perspective of the criterion-referenced-testing group. Thus, whether direct or indirect, a different line of questioning would have been necessary to elicit responses that would reveal the implicit learning theories of specialists not in the criterion-referenced-testing camp.

Other viewpoints held by this last group of testing directors, at least about the role of testing in instruction, are represented reasonably well by returning to the first level of analysis summarized in Appendix A. Respondents in Category I describe testing situations where very little instructional attention is given to tested content per se. "What's on the Iowa Test really does not determine what's going to be taught in the classroom" (I.2). And, generally, they appeared to think it was a good thing that tests do not have an undue influence on teaching. By implication, test directors in Category II also do not approve of having the test be the exclusive target for instruction because they each described mechanisms that ensure that the entire curriculum is taught, not just what's tested. Similarly, some members of Category IV and Category VI appear to reject the idea of targeting instruction by means of the test. For example, according to Test Director IV.10, "I think the issue is with teachers who are not as seasoned. For them, in particular, tests circumscribe the curriculum and determine it." Several of the respondents in Category VI, ones who did not espouse a criterion-referenced perspective, conveyed a negative tone. This last group of district test directors seems to believe that some important objectives are given short shrift because they are not tested. As noted by Director VI.27, "We do have some evidence that shows when you have a basic skills test as we do statewide that the amount of effort that goes into that does subtract from some of the higher level skills." However, none of the test directors who gave slightly negative responses about the effects of testing on instruction mentioned being concerned about basic skills testing per se or complained about the sequencing of instruction to ensure mastery of basics skills first. Rather, they seemed to be concerned that emphasis on testing had given basic skills disproportionate weight compared with unmeasured skills.

In the remaining sections of the article, I focus on the criterion-referenced model of learning held by many measurement specialists, setting aside the viewpoints of those in this last group who seem to be against measurement-driven instruction. The next section is intended to illustrate the origins of the criterion-referenced-testing perspective in behaviorist psychology. Although the third section of the article introduces a cognitive or constructivist perspective in contrast to behaviorism, there is no implication intended that these new learning theories underlie the thinking of a signifi-

cant group of measurement specialists. It seems more likely to me that this "other" group of measurement specialists holds to older views of measurement, relying on the idea of tests as samples from a broad content domain, but without a professionally shared theory of learning. (Note that traditional psychometrics was developed in the context of individual differences psychology and focused on static assessment of differences rather than the assessment of changes due to learning.)

### Origins of Measurement Specialists' Learning Theory in Programmed Instruction and Behavioral Psychology

Why is it that so many measurement specialists talk in such similar terms about the sequencing of student learning and the close alignment of tests to instruction? Several explanations are possible. It is conceivable that there is only one true way to organize effective instruction, and measurement specialists all arrived independently at the same conclusion. It is more likely, however, that measurement specialists who share similar views about learning had the same training in the educational psychology of a particular era or adopted these views implicitly when they adopted the principles of criterion-referenced testing. Most likely some combination of these explanations is at work.

My purpose here is to argue that the criterion-referenced-testing paradigm is grounded in the learning theory of behaviorism (and before that in Thorndike's connectionism) and that implicitly the majority of measurement specialists invoke this model when they think about learning. My treatment of behaviorism is necessarily simplistic, focusing on the principles that parallel those in the accounts of measurement specialists and ignoring other major aspects of the theory such as the contingencies of reinforcement. The aim is to describe what contemporary measurement specialists remember from behaviorism, not to fully elaborate the positions of the original thinkers.

Appendix B is a historical data display of quotations intended to exemplify the learning and instructional model of behavioral psychology. Whether couched in terms of teaching machines, learning hierarchies, programmed learning, mastery learning, or criterion-referenced testing, these authors share the same learning theory. This theory can be organized into two principles which correspond to the criterion-referenced-testing propositions in the first section. I summarize these principles in reverse order. Not surprisingly, the learning proposition comes first in the discourse of the psychologists, and the testing-instruction principle comes second.

1. *Learning is seen to be linear and sequential. Complex understandings can occur only by the accretion of elemental, prerequisite learnings.* In Skinner's (1954) words, "The whole process of becoming competent in any field must be divided into a very large number of very small steps, and reinforcement must be contingent upon the accomplishment of each step" (p. 94). And according to Gagne (1970), "Thus it becomes possible to 'work backward' from any given objective of learning to determine what the prerequisite learnings must be—if necessary, all the way back to chains and simple discriminations" (p. 242). The whole idea is to break desired learnings into constituent elements and to teach these one by one.

This view of learning is captured visually by pictures of learning hierarchies. For example, Glaser and Nitko (1971) diagrammed the necessary ordering of component subskills by using a series of boxes arranged linearly, as illustrated by a simple example in Figure 1. More complex hierarchies provided for several parallel sequences of prerequisites deemed essential to higher, terminal objectives. The implications of this model for instruction are conveyed best by Madeline Hunter's metaphor of a brick wall; that is, it is not possible to lay the bricks in the fifth layer until the first, second, third, and fourth layers are complete.

Given the specificity and minuteness of these analyses, one can imagine a highly complex set of instructional maps needed to address all the subject matter goals of public education. Although many prerequisite strands may be acquired in parallel, nonetheless the hierarchical and sequential nature of learning within strands is insisted upon. I might note that the image of parallel learning strands, each sequentially ordered and marked by essential milestones, is also consistent with the public's understanding of the immutability of grade level achievement, requiring grade retention as the only remedy to deficient skill acquisition (Shepard & Smith, 1989).

Perhaps the most serious consequence of the programmed learning or mastery learning model of instruction is that higher order skills, which occur later in the hierarchies, are not introduced until after prerequisite skills have been mastered. When Resnick and Resnick (in press) explained the inadequacies of associationist and behaviorist theories, they described the assumptions of *decomposability* and

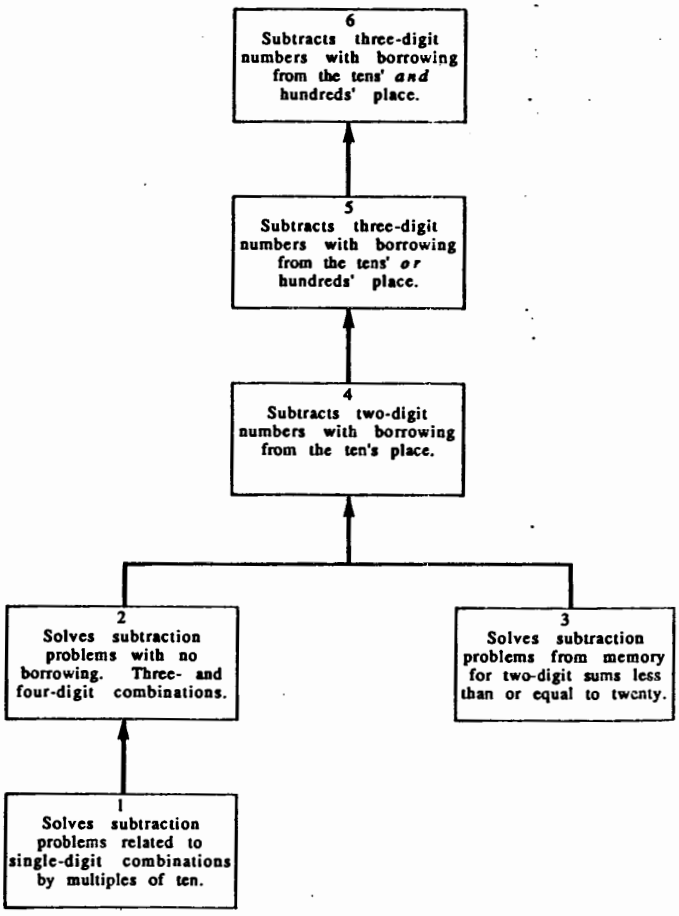


***FIGURE 1.*** **A hierarchy of objectives for an arithmetic unit in subtraction. (Adapted by permission from Ferguson, 1969, by Glaser and Nitko, 1971.)**

*decontextualization*. The model assumes that component skills can be adequately defined and mastered independently and out of context. Only then are more advanced thinking skills acquired by "adding up" or assembling component abilities.

2. *To facilitate learning, assessment should be closely allied with instruction. Tests should exactly specify desired behavioral outcomes of instruction and should be used at each learning juncture; that is, one should "test–teach–test."*

Principle 2 in the behaviorist learning model corresponds to Proposition 1 in the criterion-referenced-testing learning model implicitly held by measurement specialists. The important role of testing to judge progress in mastery learning is exemplified by several quotations in Appendix B.

> In practice, implementation of a mastery curriculum implies that children will be permitted to proceed through the curriculum at varied rates and in various styles, skipping formal instruction altogether in skills or concepts they are able to master in other ways. This demand for individualization, in turn, requires that there be some method of assessing mastery of the various objectives in the curriculum. (Resnick, Wang, & Kaplan, 1973, p. 700)

> Given our description of the learning tasks for *each* unit, we have then constructed brief diagnostic-progress tests to determine which of the unit's tasks the student has or has not mastered and what he must do to complete his unit learning. (Bloom, 1971, p. 58)

> When a student has completed a prescription, he is tested. The test is corrected immediately, and if he gets a grade of 85 percent or better he moves on to a new prescription assigned by the teacher. If he falls below 85 percent, the teacher offers a series of alternative activities to correct weakness, including special individual tutoring. He is not permitted to advance to a new unit of work until he achieves the 85 percent proficiency rating. (Education U.S.A., 1968, p. 4)

When Principles 1 and 2 are taken together, it should be clear that the behaviorist and programmed learning model also relies on assumptions about the nature of tests. First, it assumes that all important learning objectives can be specified and measured both completely and exhaustively. Each of the learning steps is small enough that highly homogeneous tests can be used to measure mastery at each step *without inference* to some broader set of test questions or criterion performances. The items for a particular objective are not thought to be sampled from a larger domain, nor is it expected that any aspect of the objective is left unassessed by the item set. If students can do what the questions ask, they have fully mastered the objective. Because each set of test items is a perfect instantiation of the learning objective, highly similar items can be used to test and retest without harm to the integrity of the measurement. It also assumed that all learning steps will be measured exhaustively at least for instructional purposes. The only circumstances where the behaviorist model admits of the need for item sampling—and therefore inference or generalizability beyond the actual test questions administered—is for review tests or placement tests, where a sampling of some of the items from some of the objectives is permitted. Even here, however, the exhaustive specification of objectives and their explicit sequencing makes the process of inference a mechanical one. It is not considered possible in this low inference system to function well on the test and not to have fully mastered the intended skills and concepts. Just as measurement specialists in the first section gave answers that treated the test and curriculum as synonymous, it should be clear from the behaviorist perspective that tests and learning objectives are equivalent and, therefore, that teaching to tested objectives is synonymous with good instruction.

These behaviorist ideas about achievement tests as perfect representations of learning goals probably took hold among measurement specialists over the last 30 years because they were compatible with concurrent trends in educational measurement that emphasized content validity rather than construct validity.[3] Construct validation was eschewed because it was the kind of "indirect" validity evidence that was needed for measurement of psychological traits. For example, Ebel (1965) argued that the validity of achievement tests should be established "directly by critical analysis of the test's specifications and contents" (p. 393) rather than indirectly by correlations with external criteria. Ebel's attention to the logical relevance of test content was a reasonable reaction against earlier statistically driven validity procedures that often distorted test content; but in my view he went too far with his assumption that it is "obvious" (p. 387) what most achievement test items measure.

Superficially, testing based on behaviorally specified objectives also seemed to be highly congruent with the evaluation models espoused by Tyler (1934) and Lindquist (1951). However, both Tyler and Lindquist had in mind much broader conceptions of learning goals and explicitly distrusted the fidelity of simplified test tasks in representing complex criterion performances. Nonetheless, the rhetoric of criterion-referenced testing or objectives-referenced testing treated all of these ideas as if they were the same. The dominance of the content-validity-only view of achievement testing held sway—despite admonitions from Cronbach (1971) and Thorndike (1971) that even "reading comprehension" is a construct requiring empirical evaluation—until the 1985 Test Standards (APA, AERA, & NCME) reaffirmed the necessity of construct validation for all types of test interpretations. The tension in the measurement literature regarding content validity may help to explain the differences between the criterion-referenced-testing specialists and the "other" group who are not willing to accept achievement tests as perfect instantiations of learning goals.

### A Competing Learning Model From Cognitive and Constructivist Psychology

But what if learning is not linear and is not acquired by assembling bits of simpler learnings? What if the process of learning is more like a Faulknerian novel where one has glimpses and a vague outline of ideas before each of the concrete elements of a story fit in place? What if learning is more like an image gradually brought into sharper focus as the learner makes connections, not stimulus–response connections but connections and relations among ideas? Or what if learning is like a mosaic with specific bits of knowledge situated within some larger design? But even these metaphors are wrong because they imply that a knowledge structure external to the student is exactly what is reproduced and cemented inside the student's head. Because we know that learning requires reorganizing and restructuring as one learns, a more organic conception is needed. In contrast to linear hierarchies, researchers now more often depict knowledge acquisition by using semantic networks that show connections in many directions. Semantic networks

can be used to capture the changing salience of concepts over time and the difference between expert and novice knowledge structures. An example from Leinhardt (1989) is shown in Figure 2.

Contemporary cognitive psychology has built on the very old idea that things are easier to learn if they make sense. We can think of learning as a process whereby students take in information, interpret it, connect it to what they already know, and, if necessary, reorganize their mental structures to accommodate new understandings. Learners construct and then reconstruct mental models that organize ideas and their interrelation. Glaser (1984) offers the following description of how instruction should be organized if we acknowledge that learning requires the greatest activity from the learner and can only be supported by good teaching.

> When schema knowledge is viewed as a set of theories, it becomes a prime target for instruction. We can view a schema as a pedagogical mental structure, one that enables learning by facilitating memory retrieval and the learner's capacity to make inferences on the basis of current knowledge. When dealing with individuals who lack adequate knowledge organization, we must provide a beginning knowledge structure. This might be accomplished either by providing overt organizational schemes or by teaching temporary models as scaffolds for new information. These temporary models, or pedagogical theories as I have called them, are regularly devised by ingenious teachers. Such structures, when they are interrogated, instantiated, or falsified, help organize new knowledge and offer a basis for problem solving that leads to the formation of more complete and expert schemata. The process of knowledge acquisition can be seen as the successive development of structures which are tested and modified or replaced in ways that facilitate learning and thinking. (p. 101)

As an example, think about learning the measurement concepts of reliability and validity. If we had a strictly linear idea about how these ideas are acquired, we might focus on mastery of prerequisite knowledge such as the standard deviation, normal curve, and the correlation coefficient. From the perspective of cognitive psychology, however, students come to the learning of these measurement concepts with a great deal of prior knowledge having to do with their own experiences in taking fair and unfair tests. Students begin with undifferentiated equivalences among good, fair, reliable, and valid tests, and ones they do well on. Good instruction is aimed at eliciting prior understandings and explicating the congruence or misfit between technical definitions and everyday conceptions. As noted by Glaser, the progression is from simpler mental models to more complex ones, rather than a progression from facts to comprehension to analysis. The first pass at textbook learning creates a mental image where reliability and validity are two equally important side-by-side constructs. Then as understanding develops, the major concepts are transformed, subordinate and superordinate concepts are recognized, hierarchies emerge, and bits of information are located in the meaning network. For example, reliability becomes subordinate to validity (see Shepard, 1990b, for an illustrative semantic net of measurement concepts).

This major principle of cognitive psychology, that learning occurs by the individual's active construction of mental schemas, applies even to the youngest children. All learning requires us to make sense of what we are trying to learn, even learning of so-called basic skills, as noted by Resnick and Resnick (in press):

> One of the most important findings of recent research on thinking is that the kinds of mental processes associated with thinking are not restricted to an advanced or "higher

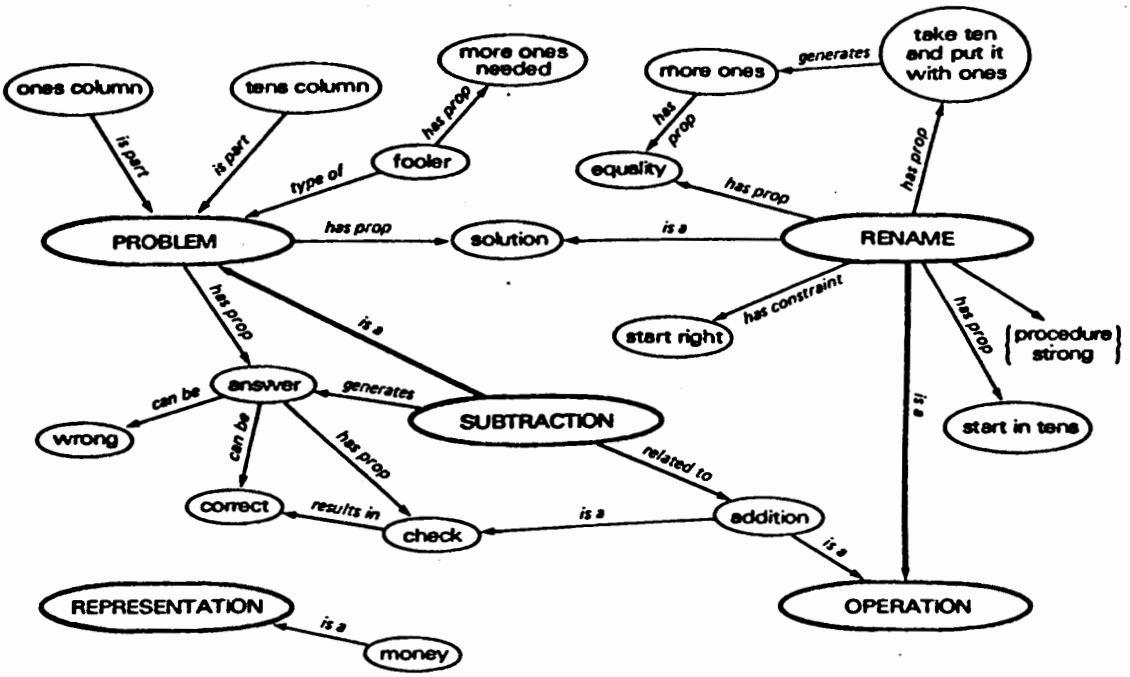


***FIGURE 2.*** **A semantic net representing one child's knowedge after a lesson on two-digit subtraction with regrouping. (Reprinted by permission from Leinhardt, 1989.)**

> order" stage of mental development. Instead, thinking and reasoning are intimately involved in successfully learning even elementary levels of reading, mathematics, and other school subjects. Cognitive research on children's learning of basic skills reveals that reading, writing, and arithmetic—the three Rs—involve important components of inference, judgment, and active mental construction. The traditional view that the basics can be taught as routine skills, with thinking and reasoning to follow later, can no longer guide our educational practice. (MS, p. 4)

The Resnicks substantiate this claim with cognitive research from both beginning reading and mathematics learning. In reading, for example, comprehension of even simple texts requires inference on the part of the reader. Authors cannot stipulate every detail needed for understanding. Competent readers supply implicit meanings and interpret the text to themselves (tell themselves the story) so automatically that they are unaware of this process until they fail to comprehend. Then good readers have strategies to reread and interrogate the text until they do comprehend. Poor readers do not engage in this kind of active translation of text that is necessary to make sense of it. Therefore, they often fail to comprehend even when they can satisfactorily decode every word.

Current research on learning has many more things to teach us about how students learn and therefore about the organization of instruction and the nature of tests that would facilitate learning. In contrasting cognitive theory with behaviorism, I have focused primarily on findings regarding cognitive structures and the notion that thinking comes before, not after, the acquisition of facts. Other fundamentally important findings have to do with the *social* aspects of learning (Resnick, 1987) and the move away from generic thinking skills to those embedded in particular knowledge domains (Glaser, 1984). To develop assessments more compatible with the cognitive view of learning would require overturning of what the Resnicks called the decomposability and decontextualization assumptions of older learning theories. Tests ought not ask for demonstration of small, discrete skills practiced in isolation. They should be more ambitious instruments aimed at detecting what mental representations students hold of important ideas and what facility students have in bringing these understandings to bear in solving new problems.

Forty years ago, Lindquist (1951) warned against the tendency to oversimplify in testing situations and recommended that tests preserve the essential complexity of criterion behaviors as they occur in natural settings. Current efforts to develop authentic or direct performance assessments (Frederiksen & Collins, 1989; Wiggins, 1989) have the same purpose, that is, to capture in test tasks the same demands for critical thinking and knowledge integration as required in desired criterion performances. Ironically, insights about how to develop and evaluate such tasks come not from the psychometric literature, nor even from laboratory experiments on cognition, but from research on learning in subject-matter fields. When learning researchers want to find out what children really know in reading, math, or science, they often interview children, ask them to think aloud while solving problems, or have them write answers to thought-provoking questions. If we want to assess children's understanding and convey to them that their reasoning is more important than a right answer, we might ask them to draw a picture to show how they set up a problem, ask them to plan a lesson for a younger child, or ask them to explain which of two proposed experiments will answer the scientist's question. Testing situations might also involve new learning, where "scores" are based on how well students are able to apply new knowledge to transfer problems.

**Conclusion: Implications for Measurement Practice**

Three main points are made in the respective sections of this article:

1. On the basis of qualitative analysis of interview data from a representative sample of 50 district testing directors, it is asserted that approximately half of all measurement specialists operate from implicit learning theories that encourage close alignment of tests with curriculum and judicious teaching of tested content.
2. These beliefs, associated with criterion-referenced testing, derive from behaviorist learning theory, which requires sequential mastery of constituent skills and behaviorally explicit testing of each learning step.
3. The sequential, facts-before-thinking model of learning is contradicted by a substantial body of evidence from cognitive psychology.

My argument is that hidden assumptions about learning should be examined precisely because they are covert. What we believe about learning and the intended effect of testing on learning should be considered directly, not "smuggled in" by the adoption of a popular test theory. What measurement specialists believe about learning does shape practice, including instructional practice. Although we have formal theories about test validity and formal means to evaluate how technical decisions affect the meaning of test scores, we do not have explicit ways to examine and debate our understandings of learning theory. Left unexamined, it is possible for a 30-year-old theory still to have a pervasive influence. Note that in selecting quotations to characterize the behaviorist position in Appendix B, I purposely chose examples from Glaser's Individually Prescribed Instruction and Resnick's earlier works. Their work in the 1980s is nearly a repudiation, certainly a significant transformation, of their earlier understandings. They have changed, but we have not, primarily because measurement specialists are no longer psychologists conversant with changes in learning theories. Thus, I propose that we engage in formal debate about our theories and expectations for the effects of tests and that we consider the empirical evidence of these effects.

The measurement community is under attack because of the negative effects of high-stakes standardized testing inaugurated by educational reform. There has been a tendency to respond defensively, as evidenced by a counterattack on performance assessment in *Education Week* (Rothman, 1990) and at a recent conference a between-sessions joke mocking authentic assessment as "measurement-free" assessment. Although I agree with some of my colleagues' fears about overly ambitious claims for authentic assessment—for example, the claim by some that performance tasks are incorruptible in high-stakes contexts—there is the danger that an entrenched technical community will be unable to respond thoughtfully to legitimate criticisms of current tests.

In the *Education Week* article, measurement specialists asserted that performance assessments are less reliable and less valid than traditional tests and that they are potentially

biased because they rely on fewer tasks; they disputed as unproved the belief that performance assessments will improve the teachng of higher order skills. Why are existing tests presumed to have the high ground in this dispute? What claim do traditional tests have to validity other than the logic of test development and actuarial correlations? Is there empirical evidence to establish the similarity in cognitive processes between multiple-choice test responses and criterion performances? What bias is introduced by asking decontextualized questions rather than having children read aloud and retell a story? If examined critically, current measurement technology rests on assumptions that are no more proved than the assertions in favor of performance assessment.

This article is an exercise in making implicit beliefs explicit so that they become available for debate and evaluation. Although differences in beliefs about learning are not the only theoretical differences or set of hidden assumptions dividing the measurement community, an understanding of learning theory is fundamental to evaluating evidence of testing effects (Is it good or bad that children spend more instructional time in drill-and-practice activities?) and therefore to framing validity investigations. If we take seriously the requirements of Messick's (1989) unified conception of validity, then new alternative assessments must be subjected to construct validity investigations designed to test the generality of conclusions about a student's knowledge and reasoning ability on the basis of what he or she is able to do in the assessment context. Judgments about face validity and correlations with teachers' grades are no longer sufficient. Furthermore, because validity includes not only the consequences of test use but also the meaning of test scores, the hypothesized effects should be systematically investigated. For example, if assessments are intended to guide instruction, then it should be demonstrable that classroom instruction for individual students is different and more effective than it would have been without the assessment information. If accountability assessments are intended to redirect instruction, then it should be possible to document whether students spend more time writing, do more extended projects, are engaged in nonalgorithmic problem solving, and so forth. These kinds of studies will be required to establish the validity of indivdidual performance assessments, not because they are more suspect than traditional tests but because these types of investigations should have been undertaken long ago to support the use of current tests.

## Appendix A

*Interview Responses of District Test Coordinators*
*Regarding Test–Curriculum Alignment and Instructional Influence of Tests (N=50)*

After responses to Questions 15–17 were read and counted as yes (Y), no (N), don't know (DK), it varies (V), or no response (NR), the question sets were reread and categorized to reflect each respondent's overall opinion about the effect of mandated tests on instruction in the district. Each category is characterized by a paraphrased summary in italic type. The number of responses in each category follows in parentheses. Categories are arranged here from least to greatest effects on instruction, *according to the respondents.* Yes, no, and don't know responses to Questions 15–17 are shown by letter abbreviations at the beginning of each quotation, for example, YNDK. Question prompts [15], [16], and [17] are shown in text to indicate which question the respondent is answering in the selected quotations. Identification codes reflecting region, size, socioeconomic status, and replicate follow each quotation. (CRT=criterion-referenced test.)

### Quantitative Summary by Question

*15. Have there been district efforts to assure that the curriculum and the district test are aligned? [aligned with the state test?]*

**No = 6**

just studying that now
What's on the Iowa does not determine what we teach.
content validity to select test but don't let it drive curriculum

**Yes = 41**

test selected to match curriculum (12)
but focus on our curriculum more (2)
but not making wholesale changes (1)
Local curriculum must reflect state test. (15)
test selected to match, then further alignment (5)
CRT test tailored to objectives (3)
customized test (2)
test driven (1)

**DK = 3**

*16. Do you think that teachers spend more time teaching the specific objectives on the test(s) than they would if the tests were not required? How much more time?*

**No = 12**

We follow our curriculum (rather than test). (5)
The test matches our curriculum. (2)
CRT, supposed to teach to objectives (1)
don't pay much attention to tests (2)
We monitor our teachers. (1)
because test samples objectives each year (1)

**Yes = 35**

definitely
always more emphasis on what's tested
We encourage them to.
because of how we give information back to them
as they get down to the wire, probably a lot more time
more than I would like
(See categorical summaries for more examples.)

**Varies = 1**
**DK = 1**
**NR = 1**

*17. To what extent do you think important objectives are given less time or emphasis because they are not included on the test?*

**None = 21**

The test reflects on our curriculum.
The test is embedded in our curriculum.
except for insecure teachers
Teachers don't worry about the test.
We monitor curriculum objectives.
teach curriculum rather than test
Teachers don't know the test yet.

**Some = 21**

There has to be a trade-off.
Yes, but these are the building-block skills.
focus on the most important objectives
has more effect on sequence, to be sure it's covered before the test
especially for unseasoned teachers

**Varies = 1**
**DK = 6**
**NR = 1**

**Examples of Responses by Category**

*I. Teachers don't worry about tests. Focus is on curriculum.* (7)

1. YNN. [16] "No, because we have our curriculum. That's the forefront. We look at the curriculum and establish our requirements based on what we feel should be taught to children. When we make our curriculum, we're looking at the state course of study. So our curriculum is closely modeled after the state course of study. [17] I think that's secondary. Maybe in some systems it becomes a primary objective, but in our system it has stayed secondary, because we feel we have a good core curriculum. We feel pleased with what the state has established as its course of study, and then our curriculum reflects that. And if it happens that that's also on the test, well and good." [3722]

2. YNN. [17] "To be honest with you, I don't think that our district or individual teachers look at the test that closely so that would not be a factor in their teaching. I would say that what's on the Iowa Test really does not determine what's going to be taught in the classroom." [4111]

3. YNN. [16] "Quite frankly, the teachers in our district don't pay a whole lot of attention to teaching to the test. They think that the test just serves a certain purpose and it only measures about 40% of what they teach anyway, so they don't worry about it. They just go ahead and teach and aren't really that worried about it." [4331]

*II. Efforts to ensure focus on curriculum, not test.* (5)

4. YNN. [16] "...They understand that it only covers a sample of the objectives in the curriculum... and they know that the objectives covered will change from year to year, and so there is not a particular way they could move other than to say we now have a testing program that really measures our curriculum; therefore, we better be sure we teach our curriculum....[17] I think there is definitely an emphasis. I mean even in test preparation, people go over test format with the kids, and the schools certainly gear up for the test. You know, they know the test is coming and we do workshops on how to sort of incorporate test taking skills and your regular instruction, not just to give item after item for kids to practice on, but have kids make up questions during the course of the year...." [1822]

5. YVN. [17] "...I'm not sure. I would guess that probably not too much. I suppose there could be some instances where that would occur, but in general, we have a curriculum for our schools set up, and they're expected to pretty much follow that curriculum. Our curriculum specialists and supervisors are out in the schools, and I would expect that wouldn't be a real problem." [2731]

6. YNN. [16] "...And I'm sure that there are individual teachers out there who might do that a few weeks before the test.... But I don't think that that is a widespread practice in the district for a couple of reasons: (1) We have an extensive teacher assessment program in the district, and it's a state required assessment program.... There is extensive observation of the teachers in the classroom. We have the essential elements that are required. Every content area has its lists of proficiencies and essential elements that are to be covered that year. There is a high level of accountability in a sense of what teachers are supposed to be doing in the classroom. Now, that's probably only going to be as good as the principals in the school, and so on, but I don't believe that this notion of teaching to the test and spending more time on these objectives is the widespread practice in the schools." [4831]

*III. Important objectives aren't slighted because test and curriculum are well matched.* (3)

7. YNN. "We have a locally developed criterion-referenced testing program, and these are skills that we have identified as being absolutely essential, and we test and retest until students show mastery. This is the kind of test that we think teachers should teach to, not particular items and answers, of course, but really focus on the curriculum, because we have identified (___) as key. In some respects, the district has put an inordinate amount of attention on achievement test results, and I can see why teachers or staff are inclined to focus on them." [1241]

8. YNN. [16] "No. I think that most of the skills that are appraised in the assessment instruments are part of our curriculum. They've always been part of the curriculum. When we're talking about skills, they've been there. I think pretty much the assessment instruments match what skills have been taught and are being taught." [1831]

9. DKYN. [16] "...For the state tests, they're supposed to teach the objectives because it is a criterion-referenced test, and the State Department of Education distributes the objectives to each and every teacher. [17] All objectives are taught." [3835]

*IV. Yes, there is an emphasis on tested objectives, but these objectives are embedded in the curriculum.* (9)

10. YYY. [17] "Yes, I do feel that there are some areas that are eliminated, not by a seasoned teacher so much, because I think a seasoned teacher who has a well-run classroom and is knowledgeable about the curriculum will teach irrespective of the test, although is aware of the test and is aware of

the objectives, but still teaches what children need to know and teaches what needs to be measured. I think the issue is with teachers who are not as seasoned. For them in particular, tests circumscribe the curriculum and determine it." [1722]

11. YYN. [16] "...I think like in any other system, once you institute a testing program, there are people who are going to look at the objectives of the test and incorporate that into their instructional program....[17] In our elementary schools, we have an instructional management system to try to ensure that teachers cover important objectives." [3831]

12. YYN. [15] "...So they established this list of essential skills. It took about a year to do that for each grade and each of those subject areas, what ought to be taught, the essential skills that ought to be taught at each grade level. And once we received these, we made sure that every teacher and administrator in our district had a copy of these, and they were instructed to make sure that they taught all of these essential skills at their particular grade level. [16] I think in our district, they probably spend a little bit more time on this, but we never did make an official correlation between our curriculum and the [state] essential skills. We never did that, purposefully in a way, because we didn't consider it worth our time, number one, and number two, we did not want to get into a situation where we put so much emphasis on this that teachers were actually being imprisoned by the state-mandated testing program and [were] either teaching the test or teaching things that were really close to what was on the test." [3241]

13. YYN. [15] "Yes. The objectives have been correlated to the curriculum. [State test?] The standardized test is the state selected test. [16] Yes. [How much more time?] I couldn't tell you that. Well, first of all, the objectives of the test are for the most part embedded in the curriculum, so they would be teaching the curriculum. But I think the emphasis is on...[what's tested]. When they get to the part of the curriculum or a skill in the curriculum that is going to be tested, then they give it more emphasis certainly, because what's tested is what's given emphasis." [3711]

*V. Yes, test focuses instruction, but these are the important objectives.* (11)

14. YYY. [16] "I think they do give added emphasis to what's on the test. In a way, we foster that feeling by making available to the teachers, I call it a 'bullet sheet,' but it is a listing that CTB offers and lists all of the 90 objectives for the test. We do push one of their reports called 'The Category Objectives Report.' It shows how well students performed on various objectives. It lays out content a little more specifically than when you just say our total reading scores, main idea, literal recall, and so forth. We push that information and use of the information. [17] You can only put so much in the 'x' amount of time the teachers have. And there are a number of tests that we administer. We give our own curriculum tests. A lot of the curriculum-based tests do have overlap on the standardized achievement test. But if you're attempting to ready kids for the achievement test, you're attempting to ready students for the curriculum tests that are developed within the local efforts. Then that could take most of the time.... But when you say less important, I don't know. The things that we try to stress are what is important. And, of course, you have terminal objectives and supporting objectives. But to push the terminal objectives which one might consider important, you have to in many respects touch upon the building block objectives." [1731]

15. YYY. [16] "Probably. They don't teach to items. We don't given them item analysis. We give them an integrated report grouped by domain. For example, for dealing with reading comprehension, we would have broken that down through a computer to facts or opinion, to main idea, to details, to sequence, to generalization. They would not see individual items. So they teach to those areas. Those areas, in turn, are curriculum referenced, and there are support materials for all of them. [17] If it's not included on the test, then we have no handle on the extent to which people pay attention to it. In the elementary [grades], the focus is basic skills, so that the focus is very much on the kinds of measures that are there which are directly related to being able to read or directly related to being able to do computations and problem solving in mathematics. I mean, it's the same as the curricula." [1811]

16. YYDK. [16] "...We do know that they are spending more time teaching those objectives, but again to clarify that, it's my feeling based on our staff development program and the sessions with those teachers involved that they are devoting more time to objectives that are measured by the tests where student performance needs to be improved." [2551]

17. YYDK. [16] "...Yes, they do, and that's particularly true because of the criterion-referenced test. For most of us, that's an intended outcome. I'm not sure it's so much more time spend on particular things as it is [that] they now organize what they present to kids in a slightly different way. They sequence instruction a little differently now because they're matching the way the course has been structured and the order in which we're going to be testing those kinds of things." [2732]

18. YYY. [16] "They would probably teach the objectives anyway, if it's part of the local curriculum. That's an interesting question. The objectives tie into the state objectives, which are supposedly measured on the state achievement tests. I know the prevailing attitude among the people in curriculum is that if the kids aren't tested on something, those teachers out there aren't going to teach it, and I don't know the extent to which that's true." [2831]

19. YYY. [15] "Well, it is a criterion-referenced test, the [state test] that I mentioned, and all of those skills are remediated, taught and then remediated after the test at every grade level, and that is its purpose, because by the time they get to be in high school prior to graduation, they must have mastered them. In order that the courts would allow us to withhold a diploma, we had to give evidence that we are teaching those skills adequately."...[16] I don't think there is any doubt...but on the other hand, I'd like to think that it is a genuine effort to improve curriculum....[17] One of the mandates in the new test committee is to find a test that does have some higher order thinking skills on it. That

is one of the things that the district is examining, and, of course, that is one of the newest developments as I see it; in all the tests now they are talking about higher order thinking skills to be incorporated in achievement tests, to give people at the top to stretch a little bit more." [3531]

20. YYN. [15] "Oh, yes! That's top priority. We have what we call the basic elements of our curriculum, and our [local tests] reflect those basic elements. [State test?] As closely as we can get it. That sometimes is a problem, but by and large, the state has made quite an effort in the last 4 or 5 years to get everybody in line for at least minimum skills or basic skills. [16]...Of course, they don't know the test items, so that they can't teach to any of the test, but they are very aware of the kinds of things that are going to be done, and so they do stress it, I'm sure. [17] I don't believe the test eliminates any really important objectives." [4832]

*VI. Tested objectives get more attention, a necessary trade-off.* (14)

21. YYY. [17] "25%. It's a trade-off." [1721]

22. YYY. [16] "If the test were not required, I don't think that anyone would spend an unusual amount of time on any objective. [17] Oh, gee, not off the top of my head, no, I can't. I guess I am generally trying to say that test from the state is extremely important to us, and if something else has to become of less importance, then so be it. That is the position that we have been put into." [2331]

23. YYY. [16] "Yes, more than I would like to see them doing, but this is true of the state test or any major test because of the emphasis that is placed on it. But you said would they still do this if the tests were not given. I think the objectives would be taught, but they might be taught in a different way . . . [17] I think we have a tendency to emphasize those objectives which are on the test. I don't think we are able to master all of those objectives that are on the test; there are some that even though they are on the test, which are not taught, and we would say that we don't expect you to teach everything that's on the CAT, but these are the things that we consider important in our curriculum that we do want you to emphasize, so it's kind of a trade-off." [2821]

24. YYY. [15] "I guess that was one of the efforts. We do change the curriculum sometimes to match the test. In other words, there are times when there's an objective being measured on an achievement test, and it might not have been included in the curriculum, and then we may add a focused area or something like that to align it a bit better. Whether that's good or not, it's done. [16] Definitely. I think more emphasis [is placed] on the local program than the state program simply because of the way we can get data back to people so that they know how to use it. [17] I think that may be true in the sense that sometimes the tests are too specific and the skills are too detailed, and then we forget the overall goal or global part of what teaching is all about. But I'm not sure if that's a problem; it probably is." [3821]

25. YYY. [15] "The state education agency now has a concern that people don't teach the essential elements; they focus on the essential elements *that are tested,* which is a narrower subset. [16] With the statewide test, yes, definitely. With our norm-referenced test somewhat, but not to the same extent. Yes, I think they do spend more time than they would if the test weren't required. [17] I don't know how to answer that in specific terms, I will give you an example. A teacher from a very upper-middle-class school, probably the highest scoring school in our district on the minimum competency test, claimed that the principal had said to them at the beginning of the year, "For this year, just forget about the curriculum and make sure the kids know the [state test] objectives." I don't know if she exaggerated, but I know that there was a lot of pressure on principals to have good scores this past year. Other principals are not as sensitive to that kind of pressure, but that's kind of a worst-case scenario. Yeah, but I think that we do leave some things out of the curriculum just because of the [press] of time." [4621]

26. YYY. [16] "I'll give you a two-part answer on that one. For the norm-referenced test, no. I do not think they spend an inordinate amount of time teaching to those objectives. I think that with the criterion-referenced test, the state-mandated test, they perhaps do in some classrooms.... There has been criticism that the test has begun to be the curriculum, and it is only minimum skills, and there is a great deal of criticism of the test for that very reason, because there is so much media emphasis and so much evaluation that is based on that, of districts as a whole, of administrators, you know, just overall, and that is one of the reasons it is being revised. [17] Well, I think if anything is, it is in those classrooms where they have concentrated on just minimum skills, finding the details, and that sort of thing. I think higher order thinking skills certainly have been excluded. There has been a great deal of emphasis, of pressure, that teachers have felt, quite frankly, to be certain that they have taught those objectives and have done it by the month that the test is given. And so to do that, they simply have made decisions to exclude certain objectives." [4711]

27. YYY. [16] "Definitely for the state and to a lesser degree for the norm-referenced test. [17] I think there's time left in the curriculum for almost all those other important objectives to be covered, and they are covered. But we do have some evidence that shows when you have a basic skills test as we do statewide that the amount of effort that goes into that does subtract from some of the higher level skills. So there is some shifting away from the higher level skills." [4741]

## Appendix B

*Quotations Exemplifying the Behaviorist Instruction and Learning Model*

**Teaching Machines** (Skinner)

"How are these reinforcements to be made contingent upon the desired behavior? There are two considerations here—the gradual elaboration of extremely complex patterns of behavior and the maintenance of the behavior in strength at each stage. The whole process of becoming competent in any

field must be divided into a very large number of very small steps, and reinforcement must be contingent upon the accomplishment of each step. This solution to the problem of creating a complex repertoire of behavior also solves the problem of maintaining the behavior in strength.... By making each successive step as small as possible, the frequency of reinforcement can be raised to a maximum, while the possibly aversive consequences of being wrong are reduced to a minimum." (Skinner, 1954, p. 94)

"Certain experimental studies of variables in programmed instruction pointedly demonstrate the importance of defined objectives to the effectiveness of the instructional enterprise. Falling in this category is the work of Gagne and his collaborators. As this method has developed, it has emphasized not only the specification of the terminal performance, but the analysis of this performance into entire hierarchies of supporting 'subordinate knowledges,' which of course are also performance objectives.

In this series of studies on various tasks of mathematics, it has been shown that the attainment of each of these 'subordinate' objectives by the learner is an event which makes a highly dependable prediction of the next highest related performance in the hierarchy. If a learner attains the objectives subordinate to a higher objective, his probability of learning the latter has been shown to be very high; if he misses one or more of the subordinate objectives, his probability of learning the higher one drops to near zero." (Skinner, 1965, pp. 29-30)

**Taxonomy of Educational Objectives** (Bloom)

"Our attempt to arrange educational behaviors from simple to complex was based on the idea that a particular simple behavior may become integrated with other equally simple behaviors to form a more complex behavior. Thus our classifications may be said to be in the form where behaviors of type A form one class, behaviors of type AB form another class, while behaviors of type ABC form still another class. If this is the real order from simple to complex, it should be related to an order of difficulty such that problems requiring behavior A alone should be answered correctly more frequently than problems requiring AB." (Bloom, 1956, p. 18)

**Programmed Instruction** (Silberman)

"This chapter includes studies which are relevant to the application of programming principles to reading instruction. The organization of this paper differs from the usual division of reading research into such topics as methods, materials, comprehension, and remediation. Instead, the following topics have been used: sequencing factors, stimulus–response factors, reinforcement factors, mediation effects, individual differences, and program evaluations. This structure corresponds with the paradigm of programmed instruction in which desired overt and covert responses are defined, stimuli are designed to evoke them, reinforcers are applied as needed, items are arranged in a systematic sequence with provision for individual differences in learning rate, and procedures are modified on the basis of learner performance." (Silberman, 1965, p. 508)

**Learning Hierarchies** (Gagne)

"The existence of capabilities within the learner that build on each other in the manner described provides the possibility of the *planning of sequences of instruction* within various content areas. If problem solving is to be done with physical science, then the scientific rules to be applied to the problem must be previously learned; if these rules in turn are to be learned, one must be sure there has been previous acquisition of relevant concepts; and so on. Thus it becomes possible to 'work backward' from any given objective of learning to determine what the prerequisite learnings must be—if necessary, all the way back to chains and simple discriminations. When such an analysis is made, the result is a kind of map of what must be learned. Within this map alternate 'routes' are available for learning, some of which may be best for one learner, some for another. But the map itself must represent all of the essential landmarks; it cannot afford to omit some essential intervening capabilities.

The importance of mapping the sequence of learnings is mainly just this: it enables one to avoid the mistakes that arise from omitting essential steps in the acquisition of knowledge of a content area." (Gagne, 1970, p. 242)

**Individually Prescribed Instruction** (Education U.S.A.)

"IPI is based on a carefully sequenced and detailed listing of 'behaviorally-stated' instructional objectives.... Each objective should tell exactly what a pupil should be able to do to exhibit his mastery of a given content and skill. This is typically something that the average student can master in one class period. Objectives involve such action verbs as solve, state, explain, list, describe, etc., rather than general terms such as understand, appreciate, know, and comprehend." (Education U.S.A., 1968, p. 6)

"When a student has completed a prescription, he is tested. The test is corrected immediately, and if he gets a grade of 85 percent or better he moves on to a new prescription assigned by the teacher. If he falls below 85 percent, the teacher offers a series of alternative activities to correct weakness, including special individual tutoring. He is not permitted to advance to a new unit of work until he achieves the 85 percent proficiency rating." (p. 4)

"IPI depends heavily on testing." Four types of tests are required: 'wide-band' placement tests to locate unit and level for each student, pretests to measure mastery of specific objectives within each unit, posttests, which are alternate forms of the pretest to determine end of unit mastery, and curriculum-embedded tests to assess within-unit progress." (pp. 11-12)

**Mastery Learning** (Bloom)

"We have used the ideas of Gagne (1965) and Bloom (1956) to analyze each unit into its constituent elements. These ranged from specific terms or facts to more complex and abstract ideas, such as concepts and principles. They even included complex processes, such as application of principles and analysis of complex theoretical statements. We have considered these elements as forming a hierarchy of learning tasks.

Given our description of the learning tasks for *each* unit, we have then constructed brief diagnostic-progress tests to determine which of the unit's tasks the student has or has not mastered and what he or she must do to complete his unit learning. The term *formative evaluation* has been borrowed from Scriven (1967) to refer to these instruments.

The formative tests are administered at the completion of each learning unit and thus help students pace their learning and put forth the necessary effort at the appropriate time.

We find that the appropriate use of the tests helps ensure the thorough mastery of each set of learning tasks before subsequent tasks are started. While the frequency of these progress tests may vary throughout the course, it is likely that more frequent formative testing may be needed for the earlier units of the course than for the later ones since typically the early units are basic and prerequisite for all subsequent units. Where the learning of some units is necessary for the learning of others, the tests should be frequent enough to ensure thorough mastery of the former units." (Bloom, 1971, p. 58)

**Hierarchically Sequenced Learning Objectives** (Resnick, Wang, and Kaplan)

"Briefly, the strategy is to develop hierarchies of learning objectives such that mastery of objectives lower in the hierarchy (simpler tasks) facilitates learning of higher objectives (more complex tasks), and ability to perform higher-level tasks reliably predicts ability to perform lower-level tasks. This involves a process of task analysis in which specific behavioral components are identified and prerequisites for each of these determined (cf. Gagne, 1962, 1968). (Resnick, Wang, & Kaplan, 1973, p. 679).

The order of objectives within each unit is based on detailed analyses of each task. These analyses are designed to reveal component and prerequisite behaviors for each terminal objective, both as a basis for sequencing the objectives and to provide suggestions for teaching a given objective to children who are experiencing difficulty. (p. 682).

In practice, implementation of a mastery curriculum implies that children will be permitted to proceed through the curriculum at varied rates and in various styles, skipping formal instruction altogether in skills or concepts they are able to master in other ways. This demand for individualization, in turn, requires that there be some method of assessing mastery of the various objectives in the curriculum....

In our classrooms, the need for assessment is met through frequent testing and systematic record keeping. A brief test for each objective in the curriculum has been written. These tests directly sample the behavior described in the objective." (p. 700)

**Criterion-Referenced Measurement** (Popham)

"In the late 1950s and early 1960s, a small but plucky band of educational innovators became entranced with the instructional potential inherent in teaching machines and programmed instruction. By transferring some powerful instructional principles, particularly those including a trial-revision teaching model, from the laboratory to the classroom in the form of a carefully sequenced or *programmed* instruction, these individuals began to achieve startling educational successes. These programmed instruction devotees would start off by explicitly defining a desired post-instruction learner behavior, build a programmed instruction sequence designed to promote learner acquisition of the behavior, then instruct and postest learners. If, in rare instances, the instruction proved sufficiently effective in its early form—yummy. But if, as was usually the case, early instructional efforts proved deficient, then the teaching sequence was revised and tried out again with new learners. Because programmed instructional sequences were essentially replicable—that is, were presented to learners by textbook or an audiovisual device in an identical fashion—such trial-revision stratgegy proved quite effective. Indeed, after a number of revisions it was quite common to secure the kind of shift in performance displayed in Figure 1-3 (a negatively skewed distribution) in which we can see that after effective instruction, the omnipresent normal curve has been bent way out of shape. After truly high-quality instruction, we find few inferior or middling performances—most learners win." (Popham, 1978, pp. 12–13)

## Notes

The research reported herein was supported in part by a grant to the Center for Research on Evaluation, Standards, and Student Testing (CRESST) from the Office of Educational Research and Improvement, Department of Education (OERI/ED). However, the opinions expressed are those of the author and do not reflect the position or policy of the OERI/ED.

[1]A reviewer raised a concern about the level of training of test directors in the sample. Although the sample was selected to be representative of the national population of district test directors, some untrained individuals might be assigned to this role and thus might invalidate my use of the data to draw inferences about beliefs in the measurement community. Unfortunately, information was not collected directly about the level of directors' training. However, two points support the conclusion that most members of the sample had technical training: (a) Large city districts were oversampled, and all attrition in the sample (from 56 to 50) occurred among small districts; and (b) the qualitative analysis did not reveal substantive differences between this sample and state directors, who would nearly always have had technical training.

[2]Intercoder reliability checks were not conducted on the accuracy of classifications because exact quantifications were not intended and because nearly all of the responses (86%) were included in the original Table 1 (Shepard, 1990b; here Appendix A); thus the reader is able to judge the credibility of the coding scheme. Although Appendix A has been substantially shortened for publication here, the original version is available on request.

[3]I am grateful to an anonymous reviewer for this insight.

## References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American School Publishers. (1989). Code of fair testing practices in education, *TestNet, 1*, 1,4.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: Handbook I, Cognitive domain*. New York: David McKay.

Bloom, B. S. (1971). Mastery learning. In J. H. Block (Ed.), *Mastery learning: Theory and practice* (pp. 47–63). New York: Holt, Rinehart, & Winston.

Bracey, G. W. (1987). Measurement-driven instruction: Catchy phrase, dangerous practice. *Phil Delta Kappan, 68*, 683–686.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average* (2nd ed.). Daniels, WV: Friends for Education.

Clark, C. M. (1988). Asking the right questions about teacher preparation: Contributions of research on teacher thinking. *Educational Researcher, 17*(2), 5–12.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice-Hall.

Education U.S.A. (1968). *Individually prescribed instruction*. Washington, DC: Author.

Ferguson, R. L. (1969). *The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction*. Unpublished doctoral dissertation, University of Pittsburgh.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Gagne, R. M. (1962). The acquisition of knowledge. *Psychology Review, 69*, 335–365.

Gagne, R. M. (1965). *The conditions of learning*. New York: Holt, Rinehart, & Winston.

Gagne, R. M. (1968). Learning hierarchies. *Educational Psychologist, 6,* 1–9.
Gagne, R. M. (1970). *The conditions of learning* (2nd ed.). New York: Holt, Rinehart, & Winston.
Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519–521.
Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist, 39,* 91–104.
Glaser, R., & Nitko, A. J. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed, 625–670). Washington, DC: American Council on Education.
Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington, DC: Author.
Leinhardt, G. (1989). Development of an expert explanation: An analysis of a sequence of subtraction lessons. In L. B. Resnick (Ed.), *Knowing, learning, and instruction* (pp. 67–124). Hillsdale, NJ: Erlbaum.
Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 119–158). Washington, DC: American Council on Education.
Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). *Comparing state and district test results to national norms: Interpretations of scoring "Above the National Average"* (CSE Tech. Rep. 308, Grant OERI-G-86-0003). Los Angeles: UCLA Center for Research on Evaluation, Standards, and Student Testing.
Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational measurement: Issues and practice, 8,* 14–22.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education, Macmillan.
Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.
Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan, 68,* 679–682.
Resnick, L. B. (1987). *Education and learning to think.* Washington, DC: National Academy Press.
Resnick, L. B., & Resnick, D. P. (in press). 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.* Boston: Kluwer Academic Publishers.
Resnick, L. B., Wang, M. C., & Kaplan, J. (1973). Task analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum. *Journal of Applied Behavior Analysis, 6,* 679–710.
Rothman, R. (1990, September 12). New tests based on performance raise questions: Assessment method said like "Star Wars." *Education Week, 10*(2), 1, 10, 12.
*Scoring High on the Iowa Test of Basic Skills, Teacher's Edition, Book B.* (1987). New York: Random House.
Scriven, M. (1967). The methodology of evaluation. In R. Stake (Ed.), *Perspectives of curriculum evaluation* (pp. 39–83). Chicago: Rand McNally.
Shepard, L. A. (1988, April). *Should instruction be measurement driven? A debate.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.
Shepard, L. A. (1990a). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice, 9,* 15–22.
Shepard, L. A. (1990b, April 17). *Psychometricians' beliefs about learning.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston.
Shepard, L. A., & Smith, M. L. (Eds.). (1989). *Flunking grades: Research and policies on retention.* London: Falmer Press.
Silberman, H. F. (1965). Reading and related verbal learning. In R. Glaser (Ed.), *Teaching machines and programmed learning: II. Data and directions* (pp. 508–545). Washington, DC: National Education Association.
Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review, 24,* 86–97.
Skinner, B. F. (1965). Reflections on a decade of teaching machines. In R. Glaser (Ed.), *Teaching machines and programmed learning: II. Data and directions* (pp. 5–20). Washington, DC: National Education Association.
Thorndike, R. L. (1971). Educational measurement for the seventies. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 3–14). Washington, DC: American Council on Eduation.
Tyler, R. W. (1934). *Constructing achievement tests.* Columbus, OH: Ohio State University.
Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46,* 41–47.