HIGHER ORDER ASSESSMENT AND INDICATORS OF LEARNING

CSE Technical Report 295

Eva L. Baker

UCLA Center for Research on Evaluation,
Standards, and Student Testing

September, 1989

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

This paper will as a chapter in *Assessing higher order thinking in mathematics* (G. Kulm, editor), to be published in 1990 by the American Association for the Advancement of Science.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

Introduction

It is now impossible to deny that testing and other approaches to the assessment of achievement represent one the most widespread and powerful approaches attempting to control the quality of schooling. It is similarly clear that the impact of tests in the service of accountability is not unbridled good. Critics contend that such tests may be shallow, they may be corruptible, or they may be incorrect (Shepard, Graue, & Sanders, 1989; Linn, 1989; Burstein, 1989; Baker, 1989; Koretz, 1989). This chapter will attempt to place in context the renewed attention to assessments which attempt to capture more complex aspects of educational attainments of students.

How is Higher Order Thinking Conceived and Measured?

Higher order thinking has been conceived both in terms of analyses of intellectual processes and of task characteristics.

Intellectual Attributes of Higher Order Thinking

In simplest terms higher order thinking measures include all intellectual tasks that call for more than information retrieval. Any transformation on information is by definition "higher order" thinking. Early writers who took the approach of detailing intellectual processes and illustrative tasks included Bloom (Bloom, 1956) and Gagne (1985). Operating from an assessment perspective, Bloom and his colleagues formulated an analysis that popularized the term "higher order," since the upper levels of their venerable *Taxonomy of Educational Objectives* (comprehension, application, analysis, synthesis, and evaluation) provided an operational description that influenced test developers and curriculum designers for years. Gagne's analysis of intellectual processes required by tasks, developed from a learning and training perspective, was of similar influence and was also frequently construed to have a hierarchical character. Both analyses rest upon the inference of transformation and construction processes from tasks, and served as important precursors to many current cognitive analyses.

Other formulations of higher order thinking derive from the general problem solving literature, and emphasize such task components as problem identification and solution testing. These may treat problem solving either as a subject-matter domain independent task, akin to general critical thinking (Ennis, 1987), or as dependent upon particular content domains. Higher order thinking can also take the form of metacognitive skills, such as planning and self-checking. These skills may be either independent of or embedded in the subject matter task to which they are applied. Other formulations focus on intellectual processes as they are unique to particular subject matter domains (e.g., the use of appropriate rhetorical structure in written composition).

Higher Order Assessment Tasks

On the assessment plane, it is common to ascribe higher order thinking to relatively global and surface features of the tasks themselves. Certain task attributes are thought to require higher order processing. For example, "open-ended " questions describe tasks for which many answers are appropriate and imply that higher order processes are necessary to respond (California State Department of Education, 1989). Obviously, open-ended questions can also solicit a range of information retrieval (e.g., tell all facts you know about dinosaurs). In the same way, almost all student constructions, such as those included in a writing or other portfolio, are assumed to be examples of higher order processes.

The concept of performance assessment is also partially connected to the measurement of higher order thinking (Baker, O'Neil, & Linn, in press). Present formulations of performance assessment involve two major dimensions: the recording of ephemeral behavior, such as electronic trouble shooting strategy or the conduct of a biology experiment, as well as the rating of resulting products or solutions. It is arguable that records of performance necessarily require higher order thinking. Consider in particular the large proportion of performance assessments in industrial or military contexts requiring the respondent to use specifically practiced operations and procedures in tasks of very narrow boundaries. One variation on performance assessment that may have a higher order component has renewed currency in public education. The approach, called "seamless" measurement, authentic assessment (Burstall, 1989) or blended assessment (Carlson, 1989), is based on the need for stronger validity in terms of both learning and instructional analysis. This approach focuses on the use of complex activities, such as experiments on the absorbancy of paper towels (Shavelson, 1989) or the preparation of analytical papers in history. (Baker & Clayton, 1989), to make judgments about achievement. Such tasks may take longer than the usual single class period to accomplish, and may be performed individually, cooperatively, or in a team environment. These activities, even when used for assessment, share many properties with good instructional lessons, including the arousal of curiosity in students, and the fact that teachers themselves may be rating the quality of students performance. In many of these examples, both student processes and products may be judged. Nonetheless, measures of higher order thinking have significant costs. Among them are the restricted number of tasks that may be sampled in a fixed period of time, the administrative and practical intricacies of reliable and accurate scoring, and the credibility of results for a society used to numbers, stanines, and norms. Alternatively, their benefits include increased validity and potential for positive curricular and instructional impact. Another important benefit may be the reduction of the salience of ceremonial testing, and its attendant costs of targeted test preparation.

Again, the avoidance of a multiple choice format or the presence of a label of open-ended or "authentic" are not sufficient to assure that these measures are actually assessing higher order processes. Details of students' actual instructional histories and the role of test administrators can alter to rote tasks measures that appear to tap the highest reaches of human thought. This reality underscores a major point of this paper: that any measurement process, higher order thinking in particular, must be understood both in the light of other available information and the intended uses of implementation.

In the next section, the multiple sources for the higher order thinking movement will be discussed. Subsequent sections will consider the policy context for measurement, including a description of the multiple indicator approach. They will also provide detailed examples of both a higher order thinking measure, and an example which might inform its understanding in a policy context. Overall, this chapter will consider the multitude of issues relevant to understanding the next wave of achievement measurement.

Impetus for the Measurement of Higher Order Thinking

The potential uses of tests condition their development. In the area of testing in general, and higher order testing in particular, there are at least three major sources for new measures: research, policy, and practice. All three of these have converged on the need of measuring higher order processes. Let us consider each in turn.

Scientific Research

The research context impacts measure development in two related ways. First, theoretically driven targets of inquiry—in other words, new constructs—help us reconceptualize our thinking about common processes. Examples of such constructs are: mental model (Norman, 1983; Collins & Gentner, 1984) advance organizer (Ausubel, 1960), "bug" (Brown & Burton, 1978), or metacognition (Meichenbaum & Asarnow, 1978). These constructs are based largely on cognitive science perspectives and provide frameworks to influence the design of human performance measures. Our community has begun to become interested in measuring problem-solving processes, in assessing group performance to capture the social meaning of certain tasks, and in measuring alternative representation modes for student knowledge, all based on scientific research. Still unresolved are the relative roles of process and product in the assessment of higher-order thinking, the place of assessment outside or within subject matter domains, and the importance of new conceptions of transfer to assessment.

Clearly the topics of research are influenced not only by what is theoretically interesting, but by what is socially important as well. Measurement foci are also similarly influenced by social goals. One way to keep score on social importance is to track the availability of research funding. For example, the 1988 Vincennes incident in the Persian Gulf where an Iranian airliner was mistakenly shot down predicted renewed research attention to measuring task performance of teams under high stress conditions, and such predictions have been verified (see, for example, U.S. Department of Defense, 1989). Any choice in R&D goals can only be judged in the light of the trade-offs in support for other goals. A case in point is the deemphasis on equity issues in the political arena during the last administration. This choice had clear consequences for acceptable assessment foci, consequences which no doubt benefitted more elitist targets, such as content-focused studies of higher cognitive performance.

A second, related influence from the domain of research derives from the actual methods used in research itself. When research approaches emphasized behavioral constructs, and quantitative methods with summary estimates of large groups held methodological currency, psychometric research and test development marched in support. Thus, there are clear implications for testing inherent in the shift during the last fifteen years to cognitive psychology. With its relegitimating of inferences from small samples (Shulman, 1986), self-report data (Ericsson & Simon, 1984) and other practices drawn from the ethnomethodology side of research (Levine, 1988), we should expect characteristics of tests to develop accordingly (e.g., limited task sampling). Collateral psychometric developments (Bock, 1987), for example, permit this transition by generating credible quantitative estimates for measures based on restrictive task sampling.

Educational Policy

The major impetus for identifying higher order thinking as an accountability target grows from policy, not science, however. Recent surveys of educational reform (Pipho, 1988) confirm that tests are used as policy instruments with increasing frequency since the publication of *A Nation At Risk* (National Commission on Excellence in Education, 1983). Tests are seen, correctly, as at least one operational way to communicate standards, and their uses have proliferated, extending beyond requirements for the award of high school diplomas to exit tests from kindergarten, for grade-to-grade promotion, merit diplomas and teacher recertification. If tests are so important, attending to their focus becomes more critical. Reasons for emphasizing higher order thinking in school programs and assessment were identified in reports by prestigious and powerful groups. Three ideas recur often in these reports.

The erosion of economic competitiveness. The international trade imbalance, the editorials about the second tier status of America, and the visible incursion of foreign wealth continue to raise sharp anxieties. These problems have been directly attributed to the failure of schools to prepare an adequately educated workforce in reports by powerful and prestigious groups such as the National Academy of Sciences, National Academy of Engineering, Institute of Medicine (1984). The comparison group of interest is obviously the Japanese, and the subject matters of concern are mathematics and science. The IEA report on US students' performance in mathematics (McKnight et al., 1987), coupled with the emphasis on formal testing in Japanese schools, strengthens for many the argument that testing will provide a way to improve student performance (National Governor's Association, 1986). One might note, with some irony, that the tests used in Japanese schools consist principally of rote items. Thus, it is probably not the test content itself that provides Japan with its educational competitive edge, but rather the relatively permanent consequences of test failure.

The failure of schools to educate disadvantaged students to a level that permits their full and productive participation in our society. Renewed awareness of the problems of disadvantaged, principally black and Hispanic, youth is fueled by media reports of increased violence, gang participation, an alarming rate of teenage pregnancies, and rampant drug traffic and use. National attention is also focused on adult literacy (see, for example, House & Madula, 1987; Kirsch & Jungeblut, 1986; Bain & Herman, 1988; Sticht, 1987; MDC, Inc., 1985).

The inability of the educational system to prepare students for change. Here, the grab bag of societal ills is attributed to failure of individuals and organizations to adapt to change. Issues such as industrial redevelopment, unsuccessful job displacement and retraining, increasing reliance on technology, and illegal immigration are lumped together (see Cohen, 1987; Carnegie Forum on Education and the Economy, Task Force on Teaching as a Profession, 1986; Goodlad, 1984; Sizer, 1984).

The indisputable fact is that these concerns stimulated a rash of specific state and local reform efforts (Bennett, 1988) and that the focus of these reforms substantially has been standards and accountability. How has educational practice reacted?

Educational Practice

Let's call educational practice the composite of what local administrators and teachers do to implement policy, to get curriculum enacted, and to teach appropriately in classrooms. Systematic educational models link practice and requirements. These models emphasize the centrality of formulating goal statements, the consequent design and implementation of curricula and instructional practices, and the subsequent creation, administration, and interpretation of measures to assess the attainment of such goals. In real life, these relationships are less rational and orderly. From a period of time when the text materials were the dominant fact of life, we have moved to a period when what is tested is of equal or more importance. At this juncture, practitioner views are predictably divided. Many critics see testing as a preemptive, coercive, implicit goal-setting device, whereas other assessment proponents see only good management. The problem used to be to understand whether the major impact of formal testing programs was to measure the impact of innovation or to reform educational programs themselves (Baker, 1989). Most policy analysts would probably now cleave to the reform function of tests.

If we believe that assessment now sets goals, we need to examine how assessments and resources are structured to allow teachers to improve their teaching.

To do so, tests should be reported as diagnostically connected to curricula, so that teachers can act upon them instructively. Knowing that a class falls into the lowest quartile is less valuable to a teacher than knowing what domains of content and skills need improvement for given children. A related aspect to the formative evaluation use of tests (Baker, 1974), is the interplay between the availability of texts and other curriculum materials and their relationship to both goals and measured outcomes. This resource issue has been a matter of major focus in some states and districts, sometimes under the label of "curriculum alignment." To align a curriculum means that attempts are made to match goals, classroom instructional resources, and tests. When the test is the only element in the alignment set with clear sanctions associated with it, the end result of alignment processes amplifies even more the curricular influence of the test. Depending upon where you sit ideologically and organizationally, this kind of efficacy may be good or evil. A curriculum gets narrowed or focused, take your pick.

For tests to impact instruction, teachers need to be able to teach higher order skills in the first place as well as to understand test results to improve their instructional strategies. Most studies of classroom practice suggest that teachers do not use many higher order teaching practices, although there is evidence that they can be instructed to do so (Pogrow, 1988). Unfortunately, the knowledge base upon which teacher training has drawn is not rich in providing clear strategies and techniques for teachers to use. Furthermore, teachers have limited management options for dealing with increasingly individualized learning requirements, and are again constrained by habit and standard school organization. A related issue is teachers' ability and interest in using test data to revise instruction. Studies of teacher's use of test information show that neither curricula nor teachers' instructional stamina could cope with the level of accuracy, detail, and frequency of information tests could provide (Dorr-Bremme & Herman, 1986). Furthermore, teachers are so used to test information that is irrelevant to the way they perceive teaching that such information is ignored without the pressure of accountability sanctions. Attempts to make higher order thinking assessments appear to be more like classroom activities and less foreign to the teaching environment should help. Furthermore, somewhat down the road, the utility of classroom databases (Herman, 1988) and enriched and well-supported computer interventions (Baker, Herman, & Gearhart, 1989) may provide needed support. But computer magic is not yet widespread. Most encouraging may be the opportunity provided by the reprofessionalization and restructuring movements. These reforms, to provide more power to teachers, more than touch on the topic of higher order thinking. Teacher organizations may very well negotiate accountability requirements in the service of focusing on important intellectual goals for children and new teaching behaviors for themselves. The recent activities on the National Board for Professional Teaching Standards (Carnegie Corporation of New York, 1989) suggests that recertification assessments will strongly focus in this direction. Thus, there is some small chance that the converging interests of science, policy and practice may actually result in serious reformation of schooling. If other strategies are selected, teachers will have the problem of delivering on policy expectations without changes in their preparation, resources, or commitment .

To sum up the points of this section, major forces have converged, promoting the higher order thinking assessment, in the research and policy communities in particular, and have raised the stakes for school accountability one more round.

Prospects for Success for Policy Driven Higher Order Thinking Assessment

What has been recent testing experience? To what extent have prior testing reforms been successful? Answers to these questions may help us to predict

if tests of higher order thinking will achieve their intended policy goals. Opinion is mixed. Case studies of testing reforms conducted in five different localities were reported by Ellwein and Glass (1987). It is their contention that testing programs possess largely symbolic value, because the educational system finds a way around standardized testing requirements. Cut scores get lowered, and other "safety nets" are strung up to protect individuals who don't succeed. Shepard, Kreitzer, and Graue (1987) in an analysis of the Texas teacher recertification test, reached similar conclusions, as did Rudner and Baker (in press) when they reviewed statewide teacher testing programs. Others are studying the concomitants of more stringent standards and testing programs. Of interest are the differential effects on minority students, effects which may increase their drop-out rates (Catterall, 1987).

How much testing of higher order thinking is going on? This question is not easily answered because of definitional problems and the aforementioned potential transformation by instruction of higher order skills into memorized procedures. In order to determine nature and distribution of higher order thinking items in mandated state programs, a study was conducted by Burstein and others (Baker, Burstein, Aschbacher, & Keesling, 1985) to document the targets of assessment. Of the states that either developed or contracted specially for state testing measures at that time, very few were found to use tests that pushed far beyond tasks of information retrieval or relabeling. Similarly, very few tests focused beyond the basic skills and assessed knowledge and skills in subject matter areas such as social studies or science. Mathematics assessment dealt largely with arithmetic. We know of specific subsequent changes in testing programs in Texas, California, Connecticut, and New York that are designed to address higher order thinking concerns, and because these states often provide leadership, we will expect to see more tests labeled "higher order" in the near future.

So higher order testing is on the way. What should we anticipate from investment in the measurement of higher order thinking skills? If one believes Ellwein and Glass (1987) higher order thinking may become little more than a symbolic flag around which to rally. One reason used to justify the development of such measures is that in time they will be able to detect the effects of reforms such as tougher curriculum standards (e.g., more time spent in class and more courses in particular subject matters). If testing results are unacceptable (i.e., not enough higher level performance is demonstrated) and political stakes are high enough, Ellwein and Glass and others believe that the "system" will find a way to blur outcomes to make them palatable to an expectant public. Such ways include changing pass scores, using less difficult items nominally to measure a higher order objective, or countenancing forms of explicit practice in instruction, so to change higher order skills functionally into retrieval behaviors. Illicit practices, such as falsifying results, also occur. At minimum, we might worry that the term higher order will be misappropriated and used to describe less challenging skills.

Some relief may come from the focus on performance based, activity structures as sources for a new regime of "standardized" tests. Allied with opportunities provided by the restructuring movement, and the leadership of influential states, the future of testing may be brighter than its history.

This entire set of events may be further perturbed by our conditioning to expect reports about achievement in the simplest terms. Fundamental misunderstandings about test norms and about the validity of single score summaries persist (Cannell, 1988a, 1988b). A mini-industry has developed in educational agencies and private firms to make test scores understandable to parents and the public, leading us to simplify, summarize, and perhaps work against the real goals we have. In no other field with a closely coupled scientific base are major results principally targeted to the least sophisticated consumer. However, the impetus of

the educational quality indicators efforts may have something to contribute to the way we conceive, develop, and report higher order thinking performance.

To summarize, a unified voice has emerged from various sectors of the community supporting the measurement of higher order skills as one way to dramatically improve American education and save us from being third-rate, or worse. But such policy uses make these testing programs vulnerable. Politicians and administrators have pledged to see to it that the schools are accountable and meet the public's expectations. Making such pledges in the dim light of what we know about teaching and learning of higher order thinking is probably both unavoidable and foolhardy. Under conditions where our instructional interventions are likely to be weak, what is left is to massage but the measure?

Assessing Higher Order Thinking: A Research-Based Example

Moving away from general sources and predictions, the next section of this chapter will illustrate an R&D project in the general area of higher order thinking.[1] Our initial task was to develop sophisticated scoring approaches to be used to assess the quality of content in student's writing. As we became immersed in the research process, we broadened our goals to include writing that was in part text based, and began to think of our task as the assessment of deep understanding. We also undertook the project with the expectation that we would extend our work horizontally, into other topic areas, vertically, to other age ranges, and diagonally to other subject fields.

Task Area

We selected the area of history and focused on the pre-Civil War era of American history. Our choice was conditioned by the paucity of good available measures in contrast to the universality of the topic in American schools. After some early textbook analyses, we decided we needed to use primary source materials as part of the stimuli for performance. Our reviews regretfully suggested that US history texts available in high school do not treat topics to any degree of depth, thus rendering the search for "deep understanding" fruitless. We selected speeches as the genre of primary material since they optimized on authenticity demands and time constraints (Baker, Freeman, & Clayton, 1989).

Thus, we are attempting to explore the construct of "deep understanding"; our present definition is tentative and relates to the following components and attendant theoretical bases:

1. Deep understanding requires the activation of thinking processes applied to specific subject matter content (i.e., history topics). These thinking processes depend upon well-known cognitive analyses of learning, including processes such as active construction in the knowledge acquisition process, and elaboration, and the integration of meaningful material into existing prior knowledge (see Brown & Campione, 1986, and the comprehensive review by Segal, Chipman, & Glaser, 1985).

2. Deep understanding may involve qualitative differences between expert and novice understanding (see Chi & Glaser, 1980). Expert understanding of topics in this area may be premise driven, allusive, and integrative, whereas novice understanding may be more literal and componential (Baker, Freeman, & Clayton, 1989).

---

[1] The author wishes to thank her colleagues on the project, particularly Marie Freeman, Serena Clayton, Yujing Li, Sheng-Chei Chang, David Neimi, Reggie Stites, and Pam Aschbacher.

3. The expression of deep understanding depends upon a sophisticated interplay among three types of knowledge: strategic, procedural, and content (or declarative). Strategic knowledge represents top-level knowledge of the major attributes and relationships of a discipline (i.e., the extent to which interpretation of events in history is context driven and the result of the interrelationships in complex factors such as politics, geography, economics, etc.; see California State Board of Education, 1988.) It also involves the understanding of the role of the historian, argument structures, verification procedures—what is often called process. We use procedural knowledge to describe routines the student uses to construct answers to our particular format of measures (e.g., how to write an essay). Content knowledge focuses on the elements inside the discipline, the principles, concepts, and facts that provide the manipulable information base. We wish further to distinguish among the contributions to student performance of prior knowledge, instruction, and the text or other text stimuli provided in the measure.

A constructivist view of comprehension suggests that understanding new material is influenced explicitly by prior knowledge and is facilitated by the activation of broad schemata, (Rummelhart, 1980; Wittrock, 1981) or to transform the language into historical terms, premises and viewpoints to provide context. Describing patterns of prior knowledge, or mental models, and their effects of comprehension of new material provides another research base against which to assess our progress (See Kieras, 1988; deKleer & Brown, 1983; Brown & VanLehn, 1980; Carpenter, Moser, & Romberg, 1982.)

Our study was initially designed to expand the content quality scoring rubric for essays in subject matter beyond those that were commonly used in written composition, (Baker, Freeman, & Clayton, 1989). Our efforts at the outset were to attempt to identify the attributes against which student essays (or longer research papers) might be judged. To this end, essays were collected after 11th grade students read either a Lincoln or Douglas speech. These were scored by two groups of experts. First, teachers trained to use an essay scoring rubric principally focused on "English teacher" issues (i.e., organization, style) scored the essays; second, a group of history teachers who were asked to score only the content knowledge exhibited in the essays and then to isolate the attributes of best and worst essays. Our data showed a remarkable degree of agreement between the two groups of raters, suggesting that matters of expression were swamping the detection of content knowledge. We also had collected think aloud protocols from teachers, historians and students who were asked to read the speeches and respond to the essay question. Our analyses of these essays suggested that experts relied heavily on prior knowledge, usually wrote from a premise or specific organizing principle, and used text information for illustration to construct their arguments. Students and some teachers, on the other hand, attended much more specifically to the presented text, sacrificed coherent argument for comprehensiveness of detail, and prepared essays that were less premise driven. After a series of pilot studies, we have developed a content scoring rubric were incorporates the following elements:

- use of prior knowledge
- principles
- facts and events
- problem/premise driven
- text information
- interrelationships
- misconceptions

In addition, an overall impression score for content quality and for essay quality is solicited.

The evolution of our scoring scheme required some revisions in the prompt solicitation. Our most recent study required two full classroom periods to complete our tasks. They included the use of an associational prior knowledge measure, a multiple-choice literal comprehension test based on the text, and task descriptions that ask explicitly for the integration of prior principles and facts with the text material. Our findings have been extremely encouraging, in terms of both construct validity (teacher judgments, standardized test scores, and transcript information) and utility and reliability of the scoring rubric. We are most interested to see the robustness of the concepts in our scoring rubric across other essay topics (i.e., the Constitution) in extended tasks, such as research papers, and in other subject areas (e.g., reports of laboratory experiments in biology).

Our immediate research plans call for the validation of this measure, and its utility at different age levels. We are also refining general specifications for measures of prior knowledge. We also wish to test the generalizability of our scoring dimensions across topics.

Our hope is to develop a scale that will assess multiple aspects of higher order cognitive performance, and that can be adapted to particular contexts. For instance, the scale might be specially weighted to emphasize the incorporation of new material in an essay, if knowledge acquisition or learning to learn were the major focus. On the other hand, if the measure were used to assess general history knowledge, more emphasis would be placed on organizing premises, prior knowledge and misconceptions. We are also in the process of exploring the use of hypertext as a way of directly assessing students' representation of content knowledge in essay planning.

Our approach to deep understanding was selected for practical practice as well as theoretical reasons: (a) extended written responses provide opportunity to observe more directly the products of complex thinking processes, (b) writing in history has strong policy impetus and potential for use, and (c) CRESST staff have experience in the qualitative rating and psychometrics of essays (Baker, Freeman, & Clayton, 1989). Our initial efforts were focused on developing stimulus materials that would: (a) meet standards of credibility for historians and history teachers, and (b) simulate instructional exposure (i.e., the long passages to be read).

Creating new measures such as our history task is a valuable undertaking, particularly when the scoring scheme for the task results in economies and flexibility. Task validity, mapped on cognitive, subject matter, and psychometric constructs, is a necessary feature but not sufficient to guarantee utility in policy contexts. Findings from measures need to be presented in a sensible way, ideally so that the decision maker understands not only the level of attainment but also some explanations for reported results. One approach to provide this context has emerged in the guise of educational quality indicator systems.

Understanding Outcomes: Educational Indicators

The metaphor of indicator systems has been adapted from the field of economics (Murnane & Raizen, 1987; Baker & Herman, 1985). Indicator systems combine data in relatively simplified models to improve understanding of the phenomena at hand. An indicator system can help us systematize the kinds of information we need to conceptualize and to condition our interpretations of test results. Although many writers have attempted to describe indicator systems (see Oakes, 1986, for a fundamental treatment of the topic in education), indicator systems typically include measures of inputs, processes, and outputs. Inputs consist of the characteristics of students, the socioeconomics of the neighborhood, and the characteristics of teachers. Processes consist of curricula and instructional

characteristics, such as courses offered and specific resources available and used. Outputs involve measures of system success, such as our history higher order thinking measure, achievement test scores, dropout rates, admission and persistence in higher education, and measured attitudes. In economics, comparable information about housing starts, money supply, levels of inventory, stock prices, employment figures, etc. provide the data. Our focus will be principally outcome measurement.

Characteristics of Indicator Systems

What are some attributes of indicator systems that make them interesting to apply to the problems in the measurement of higher order thinking?

An indicator may be a composite of many outcome measures. One articulated goal (see Baker, Linn, & Herman, 1985) for the educational achievement measurement agenda in this country is to expand the bandwidth of indicators used to make judgments about educational quality. This means that more than multiple approaches and particular measures can be used to construct an indicator. For instance, instead of focusing on a particular standardized achievement test battery as the major outcome measure for educational programs, a set of indicators would be developed, each the composite of multiple measures. In the area of higher order thinking, especially where there is some disagreement and a weak knowledge base, multiple measures could be combined, for instance, measures of problem identification, structured problem solving, decision making, inductive thinking, planning, and other forms of reasoning (see Arter & Salmon, 1987, for a more complete list). Of course, student products could be assessed as well in portfolios of student best efforts, or as benchmark measures of "open-ended" responses. Research and resource plans will determine how much triangulated or repeated measurement is appropriate. But the existence of multiple measures has benefits beyond increasing validity. When more than a single test is used as an outcome measure, the potential corruptibility of the measure is reduced. The likelihood of illicit test preparation is rendered more difficult. An indicator model also supports the idea of the interactions of individual differences and specific approaches to measurement. And its use emphasizes the multiple perspectives of ideas in this field rather than the monolithic surety communicated by hoary achievement tests.

Indicators are reported in context. Appropriate reporting of any indicator deals concurrently with other available data. Achievement indicators would be reported in the context of other input, process, and outcome indicators. These indicators would be similarly complex, consisting, as appropriate, of measures such as allocations of effort (e.g., changes in enrollment in various courses), student mobility and dropouts, teacher mobility, affective outcomes, per capita support, and class size in the school. The challenge here is to develop indicators of student educational experience to aid in the interpretation of student performance. Although the use of multiple measures is not new and has been a characteristic of many program evaluations, indicator systems are designed to be broadly institutionalized and to be robust across site and program differences.

Indicators derive their meaning over time. Indicators would provide a longitudinal and general estimate of the health of the system—what assessment programs were supposed to do in the first place, rather than posing, as some testing programs are forced to do, as finely honed measures of specific performances of the educational system. The meaning of such an indicator would be derived from changes over time (is it going up or down?) rather than from its absolute value. Also, changes in the make-up of indicators (e.g., the number of integrative, higher order thinking tasks) provide another longitudinal measure of educational quality.

Indicators foster a more participatory educational environment. There should be reduced incentive to fool with any of the particular measures used to

create the composite since any one measure would be less likely to affect the outcome value. Therefore, specifications and comparable forms for individual measures making up any composite indicator could be widely available. One feature of an ideal system might be the opportunity for local schools, or even classrooms, of teachers and of students to record their particular instructional emphases, both to describe the experiences of students and to help interpret other data. The benefit of indicators is that the summarized, combined information would still be relatively easy to understand for policy and public consumers. Schools would be encouraged to use particular measures as professional aids to the improvement of instruction, in a hypothesis-generation pattern designed to support the intellectual involvement and interest of teachers.

Context for Higher Order Thinking: An Indicators Example

Performance of outcome indicators, such as measures of higher order thinking, are among those that clearly need explanatory contexts. For example, in the recent report of a pilot study of students' performance on open-ended questions in mathematics (California State Department of Education 1989), fewer than half of the students achieved a satisfactory level, and a relatively small percentage produced competent answers. An explanation or performance is that few such problem types are taught in schools. Similarly, poor student performance can be partially explained by curriculum indicators such as lack of coverage of topics in schools, inadequacies of textbooks, low student enrollments in relevant courses, and so on. A beginning in this area is to explore the development and validation of indicators for use in state systems, and a team of UCLA/RAND researchers is engaged in developing such curriculum indicators in mathematics and in history in secondary schools. One purpose of this project is to develop interim policy indicators to assess the impact of reforms in the area of curriculum standards. Rather than waiting for outcome indicators to show improvement, curriculum indicators are designed to detect the extent to which reform intentions have found their way into enacted curricula. As such, these indicators would provide powerful explanations for subsequent performance levels.

A second goal was to provide a more comprehensive picture of the content class work and of students' course taking patterns. At minimum, these measures should respond to policy changes in course requirements.

The complexity of this project involved issues of how to conceptualize the data, how to validate any new indicator we developed, and how to develop it in an institutional form, so that regular data could be acquired without undue burden on the responders. Determining the effects of standards on course taking seems relatively straightforward. For instance, it is common to report data based on the number of students enrolled in courses by title (e.g., Algebra I or American History). A less obvious problem is that the content of a course may be vastly different depending upon who teaches it. Even where attempts are made to standardized course content in a particular district, conventions for what content is included in a given course will differ from site to site. Furthermore, some school districts create courses with even less standardized content (e.g., business math). We hoped to develop an approach that could be used at many sites and to provide a standard framework so that school, district, and state data on course content might be compared.

At the outset, we wished to determine the intent behind the institution of higher course work requirements, and this task was accomplished by a series of focused interviews (Catterall, 1988) with governors' aides, policy makers, and education leaders. In general, these interviews supported our assumptions that more stringent course requirements were instituted to improve the quality of student performance. McDonnell (1988) reported that policy makers described the

purpose of course work reforms both in general terms, "...kid's potential not being tapped...," and as more operational goals, "...to raise test scores..." (p. 5).

Our second task was to attempt to find new, comprehensive ways to determine the content of particular courses. Most particularly, we were interested in course content level of difficulty. Our data collection involved five types of data: course enrollment data from school rosters, teacher surveys, student surveys, student transcript analyses, and course materials analyses. McDonnell has reported on the utility of such data, its reliability and validity, and how feasible it is to collect and use (McDonnell, Koretz, Catterall, Burstein, & Baker, 1989). But to provide a flavor of the work, consider the issue of what it means to have one more more course required in a mathematics sequence, intended to increase students' learning of additional mathematics content. However, it is possible that content taught in a Math 1 and Math 2 sequence are simply stretched into a Math 3 course as well. Instruction is slowed down. Does this meet the policy intent? Probably not at first blush, although it is possible that students' overall performance may increase because they have a more extended opportunity to learn a fixed amount of content. To obtain measures of what was actually covered in classes, our project surveyed teachers and students. Another approach involved acquiring samples of student assignments completed by the end of a course and selecting average as well as excellent student work, a *post hoc* portfolio. Conducting topic analyses of texts and asking teachers to show us "how far" they covered provided another measure.

Needed analyses are underway to combine findings into composites that either have content validity or can be shown to have construct validity. With curriculum indicators of this sort, we should be able to answer questions about the effects of requirements on students. Do test scores rise because low performing students have become discouraged with higher standards and have dropped out? Do test scores rise because old content is being learned better over longer periods? Do test scores rise because students are mastering previously unencountered challenging content? Are there short term dips, as suggested by Koretz (1988) because less able students are taking harder classes ? Do we see a predicted drop in average scores and an increase in the variances of students taking those courses? Developing information of this sort, although a complex, time consuming task, once institutionalized, can greatly contribute to our understanding of all sorts of student performance.

Summary

This report explored the definition and impetus for the measurement attention to higher order thinking skills. Through a detailed description, a model development process was presented. This process relied on strong theory, but was firmly grounded as well in concerns for subject matter validity, feasibility, and credibility. This example led to a discussion of educational indicators as an approach to provide context for results of new measures. A brief example of the development of new curriculum indicators was included to demonstrate the complexity and utility of such efforts.

References

Arter, J.A., & Salmon, J.R. (1987). *Assessing higher order thinking skills.* Portland, OR: Northwest Regional Educational Laboratory.

Ausubel, D.P. (1960). The use of advance organizers in the learning and retention of meaningful material. *Journal of Educational Psychology, 51,* 267-272.

Bain, J., & Herman, J.L. (Eds.) (1988). *Making schools work for underachieving minority students: Next steps for research policy and practice.* Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L. (1974). Evaluation perspectives and procedures. In W.J. Popham (Ed.), *Evaluation in education.* Berkeley, CA: McCutchan Publishing Corp.

Baker, E.L. (1987, March). [Personal communication with Robert Stake, Professor of Education, University of Illinois at Champaign.]

Baker, E.L. (1987). *Time to write: U.S. Study of Written Composition.* Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L. (1989). Mandated tests: Educational reform or quality indicator? In B.R. Gifford, (Ed.), *Testing and the allocation of opportunity.* Boston: Kluwer Academic Publishers.

Baker, E.L. (1989, March). *What's the use? Standardized tests and educational policy.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Baker, E.L. (in press). Measuring deep understanding in history. In E.L. Baker & M.C. Wittrock (Eds.), *Testing and cognition.* Englewood Cliffs, NJ: Prentice-Hall.

Baker, E.L., & Clayton, S. (1989, June). *The relationship of test anxiety and measures of deep comprehension in history.* Presentation at the Conference of the Society for Test Anxiety Research, Amsterdam, The Netherlands.

Baker, E.L., & Herman, J. (1985.) Educational evaluation: Emergent needs for research. *Evaluation Comment, 7(*2), 1-12.

Baker, E.L., Freeman, M., & Clayton, S. (1989). *The measurement of deep understanding* (Report to OERI, Grant G-89-0003). Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L., Herman, J., & Gearhart, M. (1989, March). *The ACOT report card: Affects on complex performance and attitude.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Baker, E.L., Herman, J., & Gearhart, M. (1989). *The Apple Classroom of Tomorrow: 1988 evaluation study* (Report to Apple Computer). Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L., Herman, J., & Yeh, J. (1981). Fun and games: Their relationship to basic skills in elementary school children. *Educational Research Journal, 18*(1), 82-83.

Baker, E.L., Burstein, L., Aschbacher, P., & Keesling, W. (1985). *Using state test data for national indicators of educational quality: A feasibility study* (Final report to the National Institute of Education). Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L., Linn, R.L., & Herman, J.L. (1985). *Institutional grant proposal* (Proposal to the National Institute of Education for the Center on Student Testing, Evaluation, and Standards). Los Angeles: UCLA Center for the Study of Evaluation.

Baker, E.L., O'Neil, H.F., & Linn, R.L. (in press). Performance assessment framework. In S.J. Andriole (Ed.), *Advanced technologies for command and control systems engineering.* Fairfax, VA: AFCEA International Press.

Bennett, W.J. (1988). *American education: Making it work.* Washington, DC: U.S. Department of Education.

Berlin, G., & Sum, A. (1988). *Toward a more perfect union: Basic skills, poor families, and our economic future.* New York: Ford Foundation.

Bloom, B.S. (1956). *Taxonomy of educational objectives: The classification of education goals. Handbook 1: Cognitive domain.* New York: Longmans, Green & Company.

Bock, R. D. (1987). Comprehensive educational assessment for the states: The Duplex Design. *Evaluation Comment, 1,* 1-15.

Brown, A.L., & Campione, J.C. (1986). Psychological theory and the study of learning disabilities. *American Psychologist, 14*(10), 1059-1068.

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

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

Burstall, C. (1989, March). *Authentic assessment.* Paper presented at the California State Department of Education Conference titled "Beyond the Bubble," San Franciscol

Burstein, L. (1989, March). *Looking behind the "average": How are states reporting test results?* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

California State Board of Education. (1988). *History-social science framework for California public schools, kindergarten through grade twelve.* Sacramento, CA: California State Department of Education.

California State Department of Education. (1989). *A question of thinking: A first look at students' performance on open-ended questions in mathematics.* Sacramento: Author.

Cannell, J.J. (1988a). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the natinoal average. *Educational Measurement: Issues and Practice, 7*(2), 5-9.

Cannell, J.J. (1988b). The Lake Wobegon effect revisited. *Educational Measurement: Issues and Practice, 7*(4), 12-15.

Carlson, D. (1989, June). *Planning for authentic assessment.* Presentation at the Seminar on Authentic Assessment, Berkeley, CA.

Carnegie Forum on Education and the Economy, Task Force on Teaching as a Profession. (1986). *A nation prepared: Teachers for the 21st century.* Washington, DC: Author.

Carpenter, T.P., Moser, J.M., & Romburg, T.A. (1982). *Addition and subtraction: A cognitive perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Catterall, J. (1987). *School reform assessment* (Report to OERI, Grant G-86-0003). Los Angeles: UCLA Center for the Study of Evaluation.

Catterall, J. (1988). *School reform assessment: Second quarterly report* (Report to OERI, Grant G-86-0003). Los Angeles: UCLA Center for the Study of Evaluation.

Chi, M.T.H., & Glaser, R. (1980). The measurement of expertise: Analysis of the development of knowledge and skill as a basis for assessing achievement. In E.L. Baker & E.S. Quellmalz (Eds.), *Educational testing and evaluation: Design, analysis, and policy.* Beverly Hills, CA: Sage Publications, Inc.

Cohen, D. (1987). Educational technology, policy and practice. *Educational Evaluation and Policy Analysis, 9*(2), 153-170.

Collins, A., & Gentner, D. (1984). *How people construct mental models* (Technical Report No. 5740). Cambridge, MA: Bolt Beranek and Newman, Inc.

deKleer, J., & Brown, J.S. (1983). Assumptions and ambiguities in mechanistic mental models. In D. Gentner & A.L. Stevens (Eds.), *Mental models.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorr-Bremme, D.W., & Herman, J.L. (1986). *Assessing student achievement: A profile of classroom practices* (CSE Monograph Series in Evaluation No. 11). Los Angeles: UCLA Center for the Study of Evaluation.

Ellwein, M.C., & Glass, G. (1987). *Standards of competence: A multi-site case study of school reform* (CSE Technical Report No. 263). Los Angeles: UCLA Center for the Study of Evaluation.

Ennis, R.H. (1987). Testing teachers' competence, including their critical thinking. In *Proceedings of the 43rd Annual Meeting of the Philosophy of Education Society.* Cambridge, MA: Philosophy of Education Society.

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

Gagne, R. (1985). *The conditions of learning* (4th edition). New York: Holt, Rinehart.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18,* 519-521.

Goodlad, J. (1984). *A place called school.* New York: McGraw Hill.

Herman, J. (1988). *Multilevel evaluation systems project: Final report* (Report to OERI, Grant G-86-0003). Los Angeles: UCLA Center for the Study of Evaluation.

Hively, W., Patterson, H.L., & Page, S.H. (1968). A universe-defined system of arithmetic achievement tests. *Journal of Educational Measurement, 5*(4), 275-290.

House, E.A., & Madula, W. (1987). *Race, gender, and jobs: Losing ground on employment.* Boulder, CO: University of Colorado at Boulder, Laboratory for Policy Studies.

Johnston, W., & Packer, A. (1987). *Workforce 2000: American work and workers in the 21st century.* Indianapolis: Hudson Institute.

Kieras, D.E. (1988). *What mental models should be taught: Choosing instructional content for complex engineered systems.* Ann Arbor: University of Michigan.

Koretz, D. (1988). *The effects of coursework reform: Steps toward a sensitive and valid system of indicators* (Draft). Santa Monica, CA: The RAND Corporation.

Koretz, D. (1989, March). [Comments at Session 30.01, Annual Meeting of the American Educational Research Association, San Francisco.]

Krisch, I.S., & Jungeblut, A. (1986). *Profiles of America's young adults.* Princeton, NJ: Educational Testing Service.

Levine, H.G. (1988, April). *Computer-intensive school environments and the reorganization of knowledge and learning: A qualitative assessment of Apple Computer's Classroom of Tomorrow.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Linn, R.L., Graue, M.E., & Sanders, N.M. (1989, March). *Comparing state and district test results to national norms: Interpretations of scoring "above the national average."* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

McDonnell, L.M. (1988). *Coursework policy in five states and its implications for indicator development* (Working paper). Santa Monica, CA: The RAND Corporation.

McDonnell, L.M., Koretz, D., Catterall, J., Burstein, L., & Baker, E.L. (1989, March). *Developing indicators of student coursework.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

McKnight, C.C., Crosswhite, F.J., Dossey, J.A., Kifer, E., Swafford, J.O., Travers, K.J., & Cooney, T.J. (1987). *The underachieveing curriculum: Assessing U.S. school mathematics from an international perspective.* Champaign, IL: Stipes Publishing Company.

MDC, Inc. (1985). *The status of excellence in education commissions: Who's looking out for at-risk youth?* Chapel Hill, NC: Author.

Meichenbaum, D., & Asarnow, J. (1978). Cognitive-behavior modification and metacognitive development: Implications for the classroom. In P. Kendall & S. Hollen (Eds.), *Cognitive-behavioral interventions: Theory, research , and procedures.* New York: Academic Press.

Murnane, R.J., & Raizen, S.A. (Eds.) (1987). *Improving indicators of the quality of science and mathematics education in grades K-12.* Washington, DC: National Academy Press.

National Academy of Sciences, National Academy of Engineering, Institute of Medicine. (1984). *High school and the changing workplace: The employer's view* (Report of the Panel on Secondary School Education for the Changing Workplace). Washington, DC: National Academy Press.

Carnegie Corporation of New York. (1989). Certifying and rewarding teaching excellence: The National Board for Professional Teaching Standards. *Carnegie Quarterly, 34*(2), 1-7.

National Commission on Excellence in Education. (1983). *A nation at risk.* Washington, DC: US Government Printing Office.

National Governors' Association. (1986). *Time for results.* Washington, DC: Author.

Norman, D.A. (1983). Some observations on mental models. IN D. Gentner & A.L. Stevens (Eds.), *Mental models.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Oakes, J. (1986). *Educational indicators: A guide for policymakers* (OPE-01). Santa Monica, CA: The RAND Corporation.

Pipho, C. (1988). Academic bankruptcy—an accountability tool? *Education Week,* February 17, 1988.

Pogrow, S. (1988). Teaching thinking to at-risk elementary students. *Education Leadership, 46,* 79-85.

Raizen, S.A. (1988). *Increasing educational productivity through improving the science curriculum* (CPRE Research Report Series RR-066). New Brunswick, NJ: Center for Policy Research in Education.

Romberg, T.A. (1988). *Changes in school mathematics: Curricular changes, instructional changes, and indicators of changes* (CPRE Research Report Series RR-007). New Brunswick, NJ: Center for Policy Research in Education.

Rudner, L.M., & Baker, E.L. (in press). *Teacher testing: Status and prospects.* Greenwich, CT: JAI press.

Rummelhart, D.E. (1980). Schemata: The building blocks of cognition. In R.J. Spiro, B.C. Bruce & W.F. Brewer (Eds.), *Theoretical issues in reading comprehension.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Segal, J.W., Chipman, S.F., & Glaser, R. (Eds.) (1985). *Thinking and learning skills: Volume 2: Research and open questions.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Shavelson, R. (1989). *Performance assessment: Technical considerations.* Presentation at the Seminar on Authentic Assessment, Berkeley, CA.

Shavelson, R., McDonnell, L., Oakes, J., & Carey, N. (1987). *Indicator systems for monitoring mathematics and science education* (Report No. R-3570-NSF). Santa Monica, CA: The RAND Corporation.

Shepard, L.A. (1989, March). *Inflated test core gains: Is it old norms or teaching the test?* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Shepard, L.A., Kreitzer, A.E., & Graue, M.E. (1987). *A case study of the Texas Teacher Test: Technical report* (Report to OERI, Grant G-86-0003). Los Angeles: UCLA Center for the Study of Evaluation.

Shulman, L.S. (1986). Paradigms and research programs in the study of teaching: A contemporary perspective. In M.C. Wittrock (Ed.), *Handbook of research on teaching.* New York: Macmillan.

Sizer, T. (1984). *Horace's compromise: The dilemma of the American high school.* Boston: Houghton Mifflin.

Sticht, T. (1987). *Literacy and human resources at work: Investing in the education of adults to improve the educability of children* (HumPRRO Professional Paper 2-83). Alexandria, VA: Human Resources Research Organization.

U.S. Department of Defense. (1989). *Defense University Research Initiative, 1989.* Washington, DC: Author.

Webb, N. (1988). *Instructional assessment project: Quarterly report* (Report to OERI, Grant G-86-0003). Los Angeles: UCLA Center for the Study of Evaluation.

White, M. (1987). *The Japanese educational challenge: A commitment to children.* New York: Free Press.

Wittrock, M.C. (1986). *Final report of the study group on testing and cognition* (Report to OERI, Grant G-86-0003). Los Angeles: UCLA Center for the Study of Evaluation.

Wittrock, M.C. (1981). Reading comprehension. In F.J. Pirozzolo & M.C. Wittrock (Eds.), *Brain, cognition, and education.* New York: Academic PRess.