TASK STRUCTURE DESIGN: BEYOND LINKAGE*

by

Eva L. Baker and Joan L. Herman

CSE Report No. 199
1983

Center For The Study of Evaluation
Graduate School of Education
University of California, Los Angeles

*The project presented or reported herein was supported in whole or in part pursuant to a grant from the National Institute of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

ABSTRACT

The role testing can play in ascertaining and improving the effects of educational programs and services is analyzed. Our point of view maintains that the connection between tests and instruction is best made integrally through an understanding of the design of learning tasks rather than through the use of techniques that attempt to join or to link the now-separate domains of instruction and testing. The context for task structures is described, and their use in developing instruction and tests is considered. The limitations of such an approach in practice are discussed and feasible approximations outlined. Finally, the research agenda in this area is broadly sketched.

Table of Contents

	Page
Introduction	1
The Dimensions of the Problem	1
Testing in the Conceptual Framework of Science	2
Conceptual Framework for Design	5
Context: The World of Schools	6
Task Structures	9
1. Task Description	9
2. Content Limits	10
3. Generalization and Transfer	11
4. Descrimination/Performance Quality	15
5. Linguistic Features	19
6. Cognitive Complexity	20
7. Format	21
Instructional Implications of Task Structure Dimensions	22
Applications of Task Structures	25
Theoretical and Applied Research Issues	28
References	30

INTRODUCTION*

The purpose of this article is to analyze the role testing can play in ascertaining and improving the effects of educational programs and services. Our point of view maintains that the connection between tests and instruction is best made integrally through an understanding of the design of learning tasks rather than through the use of techniques that attempt to join or to link the now separate domains of instruction and testing.

The focus on the design requirements of learning tasks represents a fundamentally different perspective on the test/instruction issue. This perspective is theoretically grounded in its orientation deriving from research in learning, instruction, and cognitive processing, to name but a few areas; yet, it also has numerous potential implications for practice. The context for task structures will be described, and their use in developing instruction and tests will be considered. The limitations of such an approach in practice will be discussed and feasible approximations outlined. Finally, the research agenda in this area will be broadly sketched.

The Dimensions of the Problem

Public concern about the effectiveness of schools has led to a reliance on testing and test results that is unprecedented in recent educational history (Airasian & Madeus, this issue**; Haertel & Calfee, this issue**). Tests play prominent roles in certifying competency for high school graduation, in college admissions procedures, and in conveying through the publication of test results,

---

* This article also appears in Journal of Educational Measurement, 1982, 20(2).

** Journal of Educational Measurement, 1983, 20(2).

the effectiveness of school district policies. These examples illustrate the practical and symbolic uses of tests. Test results are regarded as the "bottom line", and educators have devoted much attention to efforts to affect such scores, and thus graduate more students, place more in better colleges, and rank their district higher in test scores among other local school districts.

Since educators have accepted the validity of tests as outcome measures, they have fed the public's desire for accountability through testing, and have created a demon that needs continually to be satisfied. Yet, the goal of improving test scores is made extremely difficult by the ways in which schools are organized and staffed, by constraints on their resources, and by the trends in the society of which schools are only a part (Bank & Williams, 1982; Zucker, 1982).

Even if these factors were optimal, the problem of creating tests that are sensitive to the results of educational programs is a difficult proposition. Part of the difficulty stems from the way in which testing, as an enterprise and a research area, has developed and part from the existing scientific, conceptual framework of testing. In our view this framework is different from the conceptual framework of instructional design.

Testing in the Conceptual Framework of Science

A scientific and experimental orientation appears paramount in the psychometric view of testing. The practical uses of testing (to create a record of student performance) are subordinated in favor of the scientific use of tests (to detect individual differences or to determine effects of interventions). The focus on the latter purpose has led to a preference for the design and development of tests that best support these uses. For instance, in the Beginning Teacher Evaluation Study (BTES), items were not selected for inclusion on the dependent measure based on their fidelity to intended learning and instruction, but rather for their correlation in pilot data with the independent variables of interest (Filby and Dishaw, 1975). This procedure increases the probability that the test will find what the researchers are looking for,but does not insure that learning is assessed adequately--because the nature and definition of learning never emerges as a primary issue.

Test development from this scientific view follows a well-established, scholarly process. Tests are developed to assess a construct. Items for the test are selected or created, and data are collected on performance. Items are included, dropped, or revised in accordance with the fit the item data provide to the posited model. In this orientation, the psychometrician's job is to hypothesize a construct or trait that is assumed to be measured by a particular item set, and then to use that set to observe nature, to report reliably what exists, and to revise the test or to reformulate the construct to explain empirical facts, processes, and their relationships. The scientist-psychometrician's mind is active, but since his/her role is a descriptive one, that is, portraying how students respond to sets of items, the scientist remains outside of the action of instruction and passive with respect to creating "better" performance. The focus is on accuracy rather than on improvement. In fact, among the most serious errors a scientist can make is to perpetrate reactivity, where inadvertent effects are produced by the process of measurement itself. The measurer should make no ripples. This is one prong of passivity in world view.

A second prong of passivity derives from the notion of stability, a central thesis underlying much of science and measurement theory. The constructs to be measured are treated as stable and are described as traits or constellations of responses thought to persist in time. Classical psychometric and statistical theories have developed under the assumption that stability, regularity, and predictability are at the heart of scientific inquiries. A concommitant assumption, especially pertinent to this discussion, is that measured individual differences probably endure over intervention. Evidence for this point of view can be found in the literature on measuring change, where growth and measurement error are sometimes treated almost interchangeably (Harris, 1963). That the early uses of tests were for placement is no surprise, since labeling and grouping individuals in homogeneous clusters is a logical outgrowth of the belief in stability.

The conflict is most simply that education is an enterprise directed at producing change, yet our tests and the psychometric theory that generates and assures them are concerned with stability and description. Traditionally, measurement's role has been that of the objective, outside observer who analyzes its subject from afar. The serious integration of testing and instruction and the use of tests for instructional improvement, however, require a more active perspective that incorporates an inside view of the phenomena of interest, learning and instruction in the educational process.

Conceptual Framework for Design

As an alternative to the scientific perspective adopted by psychometricians, design methodologies reflect a different point of view. Design is a process that synthesizes practical and theoretically-grounded ideas to produce a procedure or product that changes the environment; it is an explicit problem-solving activity that generally includes notions of planning, creation, and fine-tuning. We are all familiar with design methodologies, from the most obvious aesthetic application (graphics, interior design), those which blend aesthetic and technical features (architecture), to those which emphasize the applications of scientific findings to particular problems (computer design, engineering of all sorts, medicine).

Most professional schools are committed to training at least some of their graduates in the design (as opposed to research) paradigm, although in education, as in other fields, design has somewhat less status than research activities. Because design creates things that must operate in reality rather than contend solely with the elegance of ideas, its place in the academic community is tenuous and probably survives because some design practitioners (doctors and lawyers) are perceived to be worth high levels of financial reward. Additionally, although the outputs of design may include scientifically-based processes, it is undeniable that art or craft is also demanded. Thus, design work gets labelled "atheoretical," and its status denigrated. Even status aside, because we know less well how to help people become talented designers than competent researchers, design aspects in education are often neglected.

In contrast to the researcher, the task of the designer is not descriptive. Rather, from the outset, the designer's task is to improve upon present practice. Teachers in our public schools are design practitioners, however informally they accept that role. Their task is to combine information from a variety of theoretical approaches with practical wisdom to affect the quality of education. Change is the goal of a designer and thus that orientation would seem to serve education well, and particularly the problem of connecting tests and instruction.

Even if one were to minimize the effects of these two different frameworks on the predisposition for action or reflection, or perhaps admit that there is a continuum rather than two mutually exclusive perspectives, the serious problems remain in connecting testing to instruction and promoting school-based change. These problems inhere in the realities of school operations and in the nature of effective instruction.

Context: The World of Schools

Because psychometrics is a scholarly pursuit, most of its work is conducted in settings remote from current public school experience. Psychometric researchers don't often go to schools, and when they do, they usually don't focus on instructional issues. Because of this lack of familiarity with the lives of people in schools, it is not surprising that some misperceptions seem to have occurred.

One set of misperceptions involves teachers and what it means to teach something. Another concern is the extent to which the process of curriculum development, adoption, and implementation occurs and can be counted upon to provide a common context for school events in different schools (Sirotnik, 1981). For instance, the irrationality of the system as it most frequently operates causes only an occasional lament on the part of psychometricians. Consider that school curricula, and therefore, most of formalized content, are developed outside of the schools and marketed by text publishers. The same is true for tests used in schools. Unfortunately, the coordination between test and curriculum development is nil. Thus, at the most basic level, we can show that content differs in tests and in texts (See Floden, Porter, Schmidt & Freeman 1980) and that at the grossest exposure level, students cannot be expected to perform with regard to content they have not seen. A recent study comparing district curriculum objectives in both language and mathematics with the state assessment and various standardized tests illustrates this irrationality. Some tests included only 25% of the district curriculum, and, in some, almost 50% of the test was not covered by the curriculum (Cabello, 1982).

But attention to curricular match demonstrates only very global concern for the relationship between testing and instruction. Such a concern assumes that formal curriculum is, in fact, implemented and that simple exposure to content is sufficient for students to learn; both assumptions are unfounded. In seriously coordinating testing and instruction, one immediately is confronted with the complexity of actual instruction, including, for instance, how learning takes place, the range of content presented, the context in which a set of skills should apply, and how teachers augment or circumvent existing text material to facilitate learning. One might be tempted to ignore these confounded variables and to try to bring order to the system gradually by focusing on and controlling only one set of variables at a time. Yet, when one focuses on one aspect, say, content, and attempts to match tests and curricula on that basis, one necessarily ignores other important considerations that contribute to the irrationality of the present system. The status quo, as a result, is inadvertently perpetuated . Unless integrated alternatives are pursued deliberately, the status quo with its irrational base will continue to be our only option.

We propose, therefore, to use design methodology to create an improved system for education and learning. This approach concentrates not on linking extant curricula with extant tests, but instead on the complete design of an entire system; in fact, a redesign, so that the entire system makes sense. Instead of studying testing properties in terms of existing rules of order, we propose to focus on the learning tasks of students. The system starts with the nature and definition of what is to be learned--the task structure; the characteristics of instruction and of testing then follow naturally and rationally. Our proposition is that by designing task structures, we provide a model of the features that learning should exhibit. The model causes the requirements for testing and teaching to converge since they share, by definition, critical features. Linkage becomes redundant.

Task Structures

The task structure approach is based upon the design of the learning tasks desired of the learner. This structure is specified by a series of rules or examples; the final object is to present as clearly as possible the expected set of skills with regard to specific content. Task structures integrate critical ideas in education by reserving a place for them in the structure itself. For example, the range of content over which the learner's skill is to generalize, the manner in which transfer is treated, the behavioral formats for exhibiting performance, the level of cognitive operations required, and the complexity of language are all explicitly treated in the task structure.

We will treat the features of a task structure in turn and present definitions and descriptions, theoretical connections, and examples.

Task Description

The first element in a task structure is the general description of the task. This statement may be thought as equivalent to the statement of objective in objective-referenced tests or outcome statements used to guide the development of criterion-referenced tests. Its purpose is simply to direct and circumscribe attention to a general area of content, such as geometric proofs, and to focus on the type of skill needed, such as to solve problems or to demonstrate procedures. Such statements often have served as the sole descriptor for developing criterion-referenced tests. However, in a task structure, the statement serves principally as a convenience, as a way to get into or to approach the more taxing endeavor of describing the learning requirements of a particular task.

Content Limits

Establishing the content limit of the task is an initial problem. These limits are intended to circumscribe clearly the substance or content upon which the learner is supposed to operate. While the task description describes content in very general terms, content limits make more specific the particular elements of content to be included, and thereby help to make explicit requirement for instructional exposure and opportunity to learn. Two common approaches to content limits have been advanced: definition by curriculum and definition by agreement.

The first approach responds to the real world of school exigencies and suggests that content be defined by reference to extant curricular material (Baker, 1974), such as specifying permissible test content for a reading comprehension objective to include non-fiction selections occurring in a particular 9th grade district adopted textbook. In this instance, the content is selected based simply upon the rule of potential exposure, and while opportunity to learn is an obvious criterion, the selection of substantive material has been left to the indeterminate judgment of textbook writers and publishers. Unfortunately, analyses have shown that systematic structure and features that contribute explicitly to learning are often absent in commonly used texts (Quellmalz, Herman & Snidman, 1977; Herman, Hanelin & Cone, 1977). The reading selections at a particular grade level, for instance, do not appear to follow inferrable rules of progressive linguistic, semantic, or syntactic complexity.

A second common approach to content limits involves developing an agreed upon set of boundaries which is disseminated to both teachers and test writers. The benefits of such agreement is that the probability of a fit between instruction and testing is increased simply by communication, permitting an instructionally-focused outcome system (Popham, 1981). The criticism of such an approach rests upon the arbitrariness of the content boundaries selected (Why four line paragraphs instead of five or three line in a reading comprehension task, for example). This charge of arbitrariness has been countered by appeals to the wisdom of reliance on "human judgment."

We suggest that arbitrariness of content limits may be mitigated through reliance upon relatively strong theoretical or empirical knowledge about learning (as well as on human judgment) to decide what content limits are sensible. This approach substantially supplements the rationalization of content based upon probable exposure to extant curricula, e.g., a 9th grade textbook, and the human judgment defense, and permits each element of content limits to respond to potential issues in learning. We propose using two well documented constructs related to research on learning to define content limits: 1) generalization and transfer; 2) quality of discrimination/performance.

Generalization and Transfer. The concepts of generalization and transfer help alleviate the anathema "teaching to the test". Psychometricians, perhaps because of their frequent concentration on individual items rather than on concretely related item sets, seem to worry about this problem a good deal. Not acknowledged is the fact that "teaching to the test" can occur in two forms: teaching the exact items that appear on the test and teaching the class of content, by sampling, that the test is designed to measure. The former appeals to rote learning and is usually of limited educational value; the latter demands that we have a solid notion of what learning is intended and that we focus on significant higher-level tasks. The content limits should create this notion of the class of substance to which the test behavior is meant to apply and generalize, echoing the general idea of domain-referenced testing (Hively, Patterson, & Page, 1968).

One question here is to what topic or topics is the behavior supposed to generalize? The answer to that question depends generally on practical matters, such as the length of instructional time available for the task, or more directly, to the level of specificity at which the task is to be learned, as well as inter-relationships among potential topics. For instance, a learning task related to conjugating and applying "-ar" verbs in Spanish might be defined with the expectation that the learning should generalize to all such verbs. Alternatively, the task could be expanded to all regular "-ar, -er, and -ir" verbs, requiring a longer allocation of instruction. The conjugation procedure for these Spanish verbs is essentially the same; only details change and change in predictable ways. Research on generalization and transfer would suggest that teaching students the critical features of such conjugation and providing related practice would be sufficient for them to apply the rules to any new examples of regular Spanish verbs, and thus would support the more inclusive task structure. In contrast, the coherence and research support would be more problemmatic if the content for a task in Spanish included verb conjugations, pronoun number, and sentence expansion. The probability that transfer would occur is low, precisely because the information and concepts necessary for success vary. One decides on what level of generalization and transfer one can achieve partly by relying on unfettered human judgment, but partly based upon theoretical or empirical evidence about how knowledge in a particular subject matter area is connected (Geeslin & Shavelson, 1975).

Specifying the topics over which performance is to generalize serves an additional purpose: if transfer between topics is not explicitly taught, there is little reason to believe transfer will occur (Silberman, 1964). The definition of task structures in written composition, for example, relates to the low transfer of writing ability between topics, such as "My Best Friend" and "Credit Card Use in the United States" (Quellmalz, Capell, & Chou, 1982). Students need to learn that the same strategies and skills are applicable in both cases. The explication of the range of topics over which performance is expected or desired to generalize, when communicated to those responsible for instruction, then, can itself facilitate transfer.

Since tests take considerable time from instruction, they probably should be perceived seriously and reserved for those goals that incorporate generalization and transfer rather than for memorization of specific content. Desired levels of generalization and transfer among topics should be specified based upon the level of effort in instruction and theories of content relationships derived from analyses of the discipline (in addition to the more common curricular and consensual bases described above).

A second area to be considered under the category of generalization and transfer relates to the form in which information is presented to the learner. Form is not item format, such as passage length or number of distractors, but rather the substantive features of the task, other than topic, over which the learner is expected to generalize. For example, "triangles" is one topic in the task of learning to discriminate confusable geometric figures. To ascertain success, the learner can be asked to discriminate the correct answer presented in a single form, e.g., your standard, equilateral triangle. However, if one wants to assure an understanding of a triangle that is somewhat more robust, one would provide students with correct answers that include acute or right triangles, and perhaps triangles whose vertex is not perpendicular to the margins of the page. Students might be asked to find the triangle when other salient perceptual cues, such as size and color, might interfere. Here the issue is clearly "over what cases does the learner recognize a triangle?"

Form and type of information can be illustrated in non-perceptual concept learning as well. What class of information will the learner be expected to have acquired in order to attempt the task? For example, in written composition, a question may be posed about whether the learner has sufficient knowledge about a topic to write about it. How is that information to be provided? How complex will be the form in which the information is presented? Will it be a list that the learner simply has to transpose into prose? Will he/she be expected to infer meaning from embedded and subordinated information? As a second example, consider the task of learning to identify the main idea of a prose passage, a common enough objective of reading instruction. What type of passage will be presented? Will it be one in which there is a single clear main idea? Or one in which two partially developed ideas compete with a third "main" idea for dominance? The differences in the task intellectually should be clear, and the different requirements for instruction are probably obvious.[1]

We are recommending, then, that issues of generalization and transfer be incorporated in the content limits section of a task structure specifically to address the topics over which the response is supposed to generalize and information or presentation forms over which the response should transfer. Both of these areas are to supplement the simple notion of opportunity to learn, defined either as content in required texts (Floden, et al, 1980), or time on task, i.e., time nominally allocated to particular topic (Denham & Lieberman, 1980). It is our belief that these instances are too global to relate productively to learning tasks. In addition, we believe that attending to generalization and transfer strengthens human judgment because theoretical and empirical bases are used for content selection rather than more vague appeals to authority.

Discrimination/Performance Quality. A second, general area within the content limits section of task structures focuses on the standard of performance expected of the learner. It is at this point where claims of "educational excellence" are based by defining the required quality of response. However, performance quality should not be confused with common versions of performance standards (Mager, 1961; Popham & Baker, 1968; Anderson & Faust, 1973), all generated

1 Directly related to the issue of form of information is the linguistic features of the text of the test items. However, we will separate that discussion into a later component of task structure because the theory which supports such analyses differs from the cognitive research base of the present section.

during the behavioral objectives era of the sixties and seventies. Performance quality relates to non-quantitative features of responses that illustrate the level of refinement of the response. Because student response options fall into two major categories. Let us illustrate this principle in both selected and constructed responses.

In a selected response task such as the triangle discrimination task described above, the *a priori* difficulty of the task depends upon not only the range of correct answers the learner has to identify, e.g., isosceles and acute triangles, but the fineness of discrimination required to make that identification from distractors. Very little refinement would be required to select from distractors that consisted of those in panels a or b in Figure 1. Consider, however, if the distractors consisted of those shown in panel c. An analysis of these latter response options should demonstrate that the item requires relatively fine discriminations and exhibits higher *a priori* difficulty.

Insert Figure 1 About Here

The analysis also should clearly demonstrate that choice of distractor provides diagnostic information about the class of mistake the student is making. To select the first option in panel c, the student would have to believe that open as well as closed, three-sided, straight-lined figures met the definition of triangle. For the

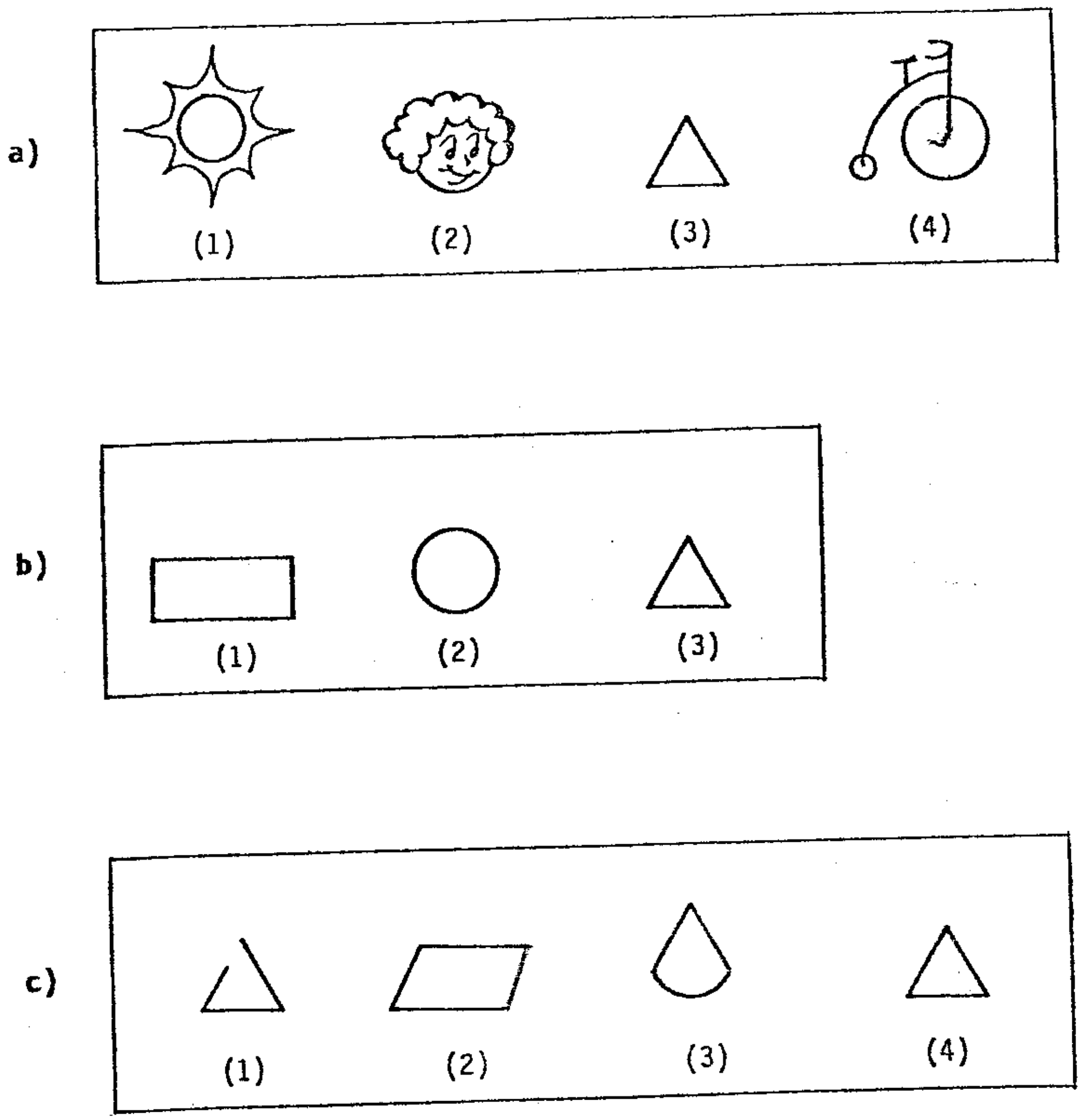

Figure 1

Alternative Distractor Options for a Triangle Discrimination Task second option, the student would have neglected the three-sided aspect of the concept. In option three, the student would have overlooked the requirement of straight sides. In each of these instances the provision of diagnostic information is preplanned.

In contrast, much of the extant literature in diagnostic testing (see Brown & Burton, 1978; Tatsuoka & Tatsuoka, 1980) lacks such *a priori* design of distractors to yield explicit diagnostic information; instead, sources of error are inferred from item response patterns. Explicitly including rules for the creation of wrong answer domains in the content limits can significantly increase diagnostic power, and to the extent that such rules incorporate research on concept learning (such as Tiemann & Markle, 1973; Tennyson, Wooley, & Merrill, 1972), the diagnostic quality will be more refined. Where concept learning is not the focus, content limits for multiple-choice items may be generated specifically to deal with aspects of the task that may have been underlearned as well as those aspects that may have been mis-learned.

In the case of constructed response, where no distractors are provided for the learner, the content limits should account for the explicit standards that will be used to judge the quality of the student effort. These standards, or criteria, are applied to student products to assess the extent to which products, such as essay answers to science questions, or English compositions, exhibit desired features. Such decisions can be reached through holistic approaches, where the overall value of the paper is judged by internal standards, or by analytic methods where particular aspects of student production, such as style, coherence or grammar, are separately considered. In either the holistic or analytic approach, the response may be judged according to a check list (where the paper, or the style, is either satisfactory or not), or through the application of a rating scale (where points from 6 to 1 depend upon the quality of student performance.)

It should be clear to see that less well explicated scoring systems, i.e., holistic, rely more on undifferentiated human judgment and experience, whereas explicated standards, such as analytic approaches with logically anchored rating scales, provide much more information about student performance. This additional information is desirable in task structures for it directly implies the type of instructional tasks the learner is expected to encounter as well as the remedies that may be necessary to address inadequate performance. Explicated standards for judging criterion responses, thus, are an important component for teaching and testing.

With the specification of content limits, performance quality is measured by design, either inherent in the level of discrimination required in selected responses or by the explicit statement of criteria in production responses. Difficulty emerges directly from the task structure design and is a function of task complexity and fineness of required discrimination rather than created empirically by proportions of people who succeed at an item. This conceptual design of difficulty may help break the tautology that exists between empirical "item difficulty" and assessment of the effects of instruction. Such an approach also allows one, by reviewing wrong answer choices, for instance, to determine when partial learning has occurred and where remediation is needed.

Linguistic Features

Linguistic features are another important aspect of task structures, but their role in test and task design has been treated in generally disjointed fashion. Level of difficulty has been assessed by various readability formulae which take into account the difficulty level of words (inferred from developmental or frequency measures), and sometimes the complexity of syntax (Duffy, 1981). Yet more complex linguistic structures play a role in tasks that either present verbal material as stimuli, including verbally stated alternative responses, or include rating systems based on verbal products by the respondent (Duffy, Curran, & Sass, 1982). Particularly when non-native English speakers are assessed, the variation in performance created by apparently casual linguistic options may be great. Bauman (1982), for example, found that problem types identified through linguistic analysis posed serious difficulties for readers--problems that were not directly related to the construct being assessed.

Systematic attention to the linguistic components of tasks may permit more accurate assessments of true performance levels. Measures of linguistic complexity need to be created that are appropriate for both long and short verbal passages and which include some notion of deviation from semantic and syntactic experiences of the respondents. For instance, some difference score may be obtained depending on the compatibility of the sentence patterns with the native language, or the root of more difficult words and the native language. To the extent that language proficiency is not an inherent feature of the task of interest, then effort should be made to purge verbal materials of unnecessary complexity.

Cognitive Complexity

Another critical feature of a task structure is the cognitive complexity of the task. Simply stated, cognitive complexity is the intellectual "level" apart from content, at which the learner is expected to perform. These levels have been taxonomized by Bloom and his associates (1956) to include six categories, Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. Presumably, each of these categories refers to cognitive processes that are successively increasingly complex as well as dependent upon prior levels. A slightly different structure has been posited by Gagne (1975) where essentially stimulus-response learning, multiple discrimination, concept learning and problem solving form the major dimensions of intellectual skills. Simplifications of these schemes have been found in the cognitive literature (Quellmalz, 1982) where principal distinctions have been made for tasks whose purpose is the storage, association, and retrieval of information contrasted with tasks requiring processing of information, including subordination, reconfiguration, and other adaptive processes.

The task structure must clearly provide an indication of the intended cognitive complexity of the task. This ascription will relate to two features of the content limits already described, generalization and required performance quality. First, complexity is a function of the degree of expected generalization and transfer, and in the nature of the required performance, e.g., the number of cues provided, the amount of information that must be subordinated should be reflected in this area. Second, complexity is also a function of the performance quality that is demanded. A good example might be a problem solving task involving the correction of a operating defect on a jet aircraft. Perhaps no transfer is necessary, for it is the F-14 and only the F-14 that is of interest. However, because of the enormous inherent complexity of the circuitry of various systems, the task requires within it a high level of discrimination and therefore, has high cognitive complexity.

Format

Another related feature of the task is the format in which the assessment is made. Format includes both the descriptive mode in which the task is presented, e.g., print, graphics, video recording, and the form in which the response is desired, e.g., multiple choice with four response options, written composition, and so on. Obviously the format relates both to the practical matter of presenting and obtaining task related information as well as to the requirement to incorporate specifications of task structure identified in content limits, in linguistic features, and in cognitive complexity. It is possible, for instance, that format is truly an unimportant issue, and that fact is demonstrated by the expectation that students will be able to demonstrate task mastery in one of many formats or in all of a number of formats.

The extent to which format dependence has taken over from optimal learning requirements of tasks is documented by the attention the general education system directs to test wiseness. Here the format of the test is regarded as separate from, and sometimes equal to (in importance), the content and intellectual skill demanded by the task itself. Including format as a particular dimension of the task structure allows for the rational review of the role of format and its relative importance or subordination to issues of content mastery. In addition, the internal consistency of descriptions about generalization of content performance quality and cognitive complexity can be assessed in reviewing the format(s) projected for task demonstration.

Instructional Implications of Task Structure Dimensions

The premise of this paper is that attention to task structure dimensions outlined above provides a common focus for and defines the structure of assessment and instructional systems. In this section, we propose to identify the aspects of the task structure that inexorably lead to instructional decisions. The problem in relating instruction and assessment changes dramatically. Instead of dealing with the amount and degree of overlap between activities and artifacts, one focuses on the degree of implementation of the task structure itself, a far different task intellectually, and with the potential, at least, for greater satisfaction.

In dealing with instruction, let us exclude from our discussion issues related to affective, motivational or social learning paradigms and focus, for purposes of our analyses, on the cognitive and behavioral tasks of learning and teaching. Clearly, based on the literature in instruction (Bower & Hilgard, 1982; Gagne & Briggs, 1981; Traub, 1966), a critical issue is the extent to which students have been exposed to a particular task and in fact have had the chance to practice it under conditions implied by the content levels, i.e., with both particular and generalized examples, and at the level of performance quality (such as discrimination), implied by the task. However, opportunity to practice criterion behavior is necessary but may not be sufficient for less able students and more complex tasks.

If criterion behavior is too complex to be acquired by repeated rehearsals, what should be done first, what component skills must be acquired? Unlike many statements of objectives, the subordinate components of instructional tasks are inherent in the task structure itself. The identification of features over which the performance is expected to transfer specifies a set of experiences for the student. For example, if the task structure is to be able to analyze particular propaganda devices in advertisements, news articles, editorials, and verbal appeals, then students would need practice with all specified media as well as instruction and practice with each specified device. The indicated embeddedness and subtlety of propaganda use would similarly suggest the successive range of difficulty that would be appropriate for instruction. In other words, inherent in the task structure is a plan for successive approximation of the end desired learning tasks, where individual components are practiced and then combined in increasingly complex sets.

The nature of instructional tasks also follows from the specification of content limits for performance quality: the classes of concepts included in the distractors, or the criteria by which the ultimate student product is to be judged. For instance, in the triangle discrimination task described above, instruction would need to take clear and differentiated account of the attributes of the triangle of interest: it is a geometric figure; it has straight lines; it is closed; it is three sided. The order in which these are treated or the motivational context in which these attributes are introduced make little difference to this analysis. The implementation issue is the extent and degree to which these attributes are treated, i.e. the extent to which instruction and practice deal with each attribute, singly and/or in combination, which represents a significantly more refined view of opportunity to learn.

Similarly, in constructed responses, if a learner's writing is to be judged on his/her use of coherent sentences in a paragraph and the choice of development used in the paragraph, then the instruction must, in a differentiated way, treat these options. Again, the context in which instruction occurs or the instructional approach, is not of first concern; matters of presentation style, sequence, etc., are not the primary focus because valid differences cannot be discerned in the absence of specified treatments, treatments which are directly relevant to and derived from the desired learning. Once more, the issue becomes whether the elements of the task structure can be found in the instructional provisions for the students. It is an implementation problem, looking at frequency and intensity, rather than a problem of determining overlap.

Metastructures for approaching these individual instructional components of the task structure depend on the educational philosophy and instructional style preference of the teacher. Direct instruction (Rosenshine, 1982) and task analyses (Gagne, 1977) approaches to teaching may be appropriate. On the other hand, a less directive, more inquiry-oriented approach may be preferred. Most important to note, however, is that the action changes from attention to the process of instruction, or how instruction occurs, to the substance of instruction and the modelling of the structure of task itself.[2] Our belief is that the way to outcomes is far easier and of secondary importance if the quality of outcome desired is sophisticated and well described. Targetted instruction under whatever approach, will likely be more effective than more diffuse attempts. You have to teach "it" if "it" is going to be learned.

Applications of Task Structures

Since it is obvious that the rhetoric of design and change is insufficient itself to create the conditions for implementation in education, what is the likelihood that such an approach is practical at all? Organizations responsible for implementing educational practices like public schools, are often not change-oriented themselves. They would rather adopt the surface appearance of change and innovation (Pincus, 1975) than to undergo the dislocation that real change implies.

Having laid out our ideas on task structures and the promise they hold for making the educational process more rational, fair, and instructionally effective, let us consider their possibilities in practice. Or have we, like the academic friends we've criticized, proposed an ivory tower system that will not survive the test of reality?

First, let us consider a serious distortion of the ideas we espouse: minimum competency testing. Essential in this movement is

2 We are indebted to Wells Hively for this part of the analysis.

the idea that schools should be responsible for assuring the acquisition of particular skills--learning tasks--and that these skills should be the subject of both instruction and testing. Yet, in practice, the target skills do not truly reflect school and teachers' main goals, and the natural linkage of instruction and testing within the system has not often occurred. Insufficient technical expertise, often volatile political environments, and high stakes have combined to produce more rather than less irrationality: ninth grade--or lower--skills many represent essentially a new one-year remedial curriculum masquerading as the minimum competency for high school graduation. Such may be the fate of most top-down change mandates that attempt to solve complex educational problems with simplistic solutions that are insensitive to local context.

Our experience, however, indicates that more positive outcomes are possible, and that approximations of our learning task approach are feasible in practice. Below we allude to two approaches we have used to implement task structures. The two examples vary in the local motivation for change and the source of educational goals--or the learning tasks to be accomplished. The examples illustrate a "minimal" and "maximum" attempt at change.

With a minimalist view, one school district attempted to solve a common district problem, "Raise those test scores." Learning tasks were directly inferred from the actual content of the tests in question, i.e. the task structures were defined to parallel state assessment test content. District curricula were analyzed to determine the extent to which instruction and practice were provided for each learning task. Little direct and explicit overlap was found and supplementary practice exercises and cues for instruction were developed to fill in the gaps. Additionally, test performance was analyzed school-by-school within the district and school specific instructional prescriptions were created. School-wide strategies and explicit instructional guidance and materials for teachers and students were designed. The entire effort was initiated centrally and received strong district leadership and subsequent principal support. While some might question the validity and value of such "teaching to test" activities, the effort was directed at instructional improvement, based on the goals measured by the test, and served the practical needs of the subject school district.

More comprehensive change efforts have been conducted in other local contexts, using a more grass roots approach. Several change efforts conducted by the UCLA Center for the Study of Evaluation have used a multiprong curriculum-assessment-staff development strategy. In these instances, teachers have been trained in the task-structure approach to integrating instruction and testing and in sound test development techniques. Teachers then play the active role, with some technical assistance, in defining critical learning tasks for their subject area, in explicating the dimensions of each task and in constructing suitable test items. The resultant tests are subsequently used to diagnose individual, class, school, and district needs, and to monitor student achievement. Model instructional approaches and teaching lessons for the target learning tasks also support the process.

More comprehensive implementation of task structures is possible. Applications in emerging technology and in the private sector in highly technical training environments represent two potential opportunities. In both cases, the incentive for high quality training may be possible to an extent not present in public school education. In the second case, there are controls on the selection of the group to receive education, either because they are hired or otherwise screened, and teaching conditions and student motivation are more tractable. The use of technology is a seductive arena not only because the personal and idiosyncratic mediation of instruction by teachers will be avoided, but also because of the possibilities for closer monitoring and immediate feedback with refined branching and remediation options.

Theoretical and Applied Research Issues

The role of theory in research on task structures is obvious. However, the theory of interest is not psychometric theory, but rather propositions that grow from perspectives in cognitive and behavioral learning in the field of psychology, in psycholinguistics, and in contrastive linguistics. A practical issue relates, once more, to the level of generality necessary and the inherent relationships among features of the task structures. For example, can one have relatively simple content and require sophisticated cognitive processes? The answer on a single instance level is "of course", but how general is that answer?. What is the relationship between language complexity, cognitive processes, and transfer and generalization of content? How circumscribed or broad can a task structure be; that is, what are the limits or optimal levels of generalization?. These and other more provocative questions need exploration as well as testing in alternative contexts and degrees of implementation. At any rate, what we hope will happen is that those with psychometric skills and those whose expertise is in the areas of learning and instruction will meet intellectually and jointly continue the task of focusing educational productivity on learning tasks.

REFERENCES

Anderson, R.C. & Faust, G.W. Educational Psychology--The Science of Instruction & Learning. New York: Harper and Row, 1973.

Baker, E.L. Beyond objectives: Domain referenced tests for evaluation and instructional improvement. Educational Technology, 1974, 14, 10-21.

Bauman, J. Linguistic Structure and the Validity of Reading Comprehension Tests (Final Report to National Institute of Education). Washington, D.C.: Center for Applied Linguistics, May 1982.

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

Bower, G.H., & Hilgard, E.R. Theories of Learning. Englewood Cliffs, New Jersey: Prentice-Hall, 1981.

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

Cabello, B. Analysis of district curriculum, CAP objectives and 3 standardized tests. Los Angeles, California: Center for the Study of Evaluation, 1982.

Denham, C., & Lieberman, A. (Eds.) Time to learn. Washington, D.C.: National Institute of Education, May 1980.

Duffy, T.M. Readability formulas: What is the use? San Diego, California: U.S. Navy Personnel Research and Development Center, November 1981.

Duffy, T.M., Curran, T.E., & Sass, D. Document design for technical job tasks: An evaluation. San Diego, California: U.S. Navy Personnel Research and Development Center, April 1982.

Filby, N., & Dishaw, M. Developments and refinements of reading and mathematics tests for grades 2 and 5. Technical Report III-1, Beginning Teacher Evaluation Study, San Francisco, California: Far West laboratory for Educational Research and Development, 1975.

Floden, R.E., Porter, A.C., Schmidt, W.H., & Freeman, D.J. Don't they all measure the same thing? Consequences of standardized test selection. In E.L. Baker and E.S. Quellmalz (Eds.), Educational testing and evaluation. Beverly Hills, California: SAGE Publications, 1980, 109-120.

Gagne, R.M. Analyses of lectures. In L. Briggs (Ed.) Instructional design: Principles and applications. Englewood Cliffs, New Jersey: Educational Technical Publications, 1977.

Gagne, R.M., & Briggs, L.J. Principals of Instructional Design. New York: Holt, Rinehart, and Winston, 1974.

Geeslin, W.E., & Shavelson, R.J. An exploratory analysis of the representation of mathtmatical structure in students' cognitive structures. American Educational Research Journal, Winter 1975, 12(1), 21-39.

Harris, C.W. (Ed.) Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963.

Herman, J., Hanelin, S., & Cone, R. Instrumentation for the Early Childhood Education Program. Paper presented at the annual meeting of the American Educational Research Association. New York, 1977.

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

Mager, R.F. Preparing instructional objectives. Palo Alto, California: Fearon Publishers, 1961.

Pincus, J. Incentives for innovation in the public schools. Review of Educational Research, 1975, 45(1).

Popham. W.J., & Baker, E.L. Rules for the development of instructional products. Inglewood, California: Southwest Regional Laboratory for Educational Research and Development (SWRL), 1968.

Popham, W.J. Modern educational measurement. Englewood Cliffs, New Jersey:Prentice-Hall, 1981, p. 211.

Quellmalz, E. Cognitive models for linking testing and evaluation. Los Angeles, California: Center for the Study of Evaluation, 1982.

Quellmalz, E., Capell, F., & Chou, C. Effects of discourse and response mode on the measurement of writing competence. Journal of Educational Measurement, 1982, 19, 241-258.

Quellmalz, E., Herman, J., & Snidman, N. Toward competency-based reading systems. Paper presented at the Annual Meeting of the American Educational Research Association. New York, 1977.

Rosenshine, B. The master teacher and master developer. Paper presented at the Annual Meeting of the American Educational Research Association. New York, 1982.

Silberman, H.F. Experimental analysis of a beginning reading skill. Santa Monica, California: Systems Development Corporation, 1964.

Sirotnik, K.A. The contextual correlates of the relative expenditures of classroom time on instruction and behavior: An exploratory study of secondary schools and classes (A Study of Schooling Tech. Rep. No. 26). Los Angeles, California: University of California, Laboratory in School and Community Education, 1981.

Tatsuoka, K., & Tatsuoka, M.M. Detection of aberrant response patterns and their effects on dimensionality (Research Rep. 80-4). Urbana, Illinois: University of illinois, Computer-Based Education Research Laboratory, 1980.

Tennyson, R.D., Wooley, F.R., & Merrill, M.D. Exemplar and non-exemplar variables which produce correct classification errors. Journal of Educational Psychology, 1972, 63, 144-152.

Tiemman, P., & Markle, S.M. Remodeling a model: An elaborated hierarchy of types of learning. Educational Psychologist, 1973, 10, 147-158.

Traub, R.E. Importance of problem heterogeneity to programmed instruction. Journal of Educational Psychology, 1966, 57, 54-60.

Williams, R., & Bank, A. Use of data to improve instruction in local school districts: Problems and possibilities. In C. Aslanian (Ed.), Improving Educational Methods: Impact on Policy. Beverly Hills, California: SAGE Publications, 1981.

Zucker, L.G. Institutional structure and organizational processes: The role of evaluation units in schools. In A. Bank & R. Williams (Eds.), Evaluation in school districts: Organizational perspectives, CSE Monograph Series in Evaluation, 1981, 10, 69-90.