VOLUME 27 NUMBER 7 MARCH 1995

FOCUS ON EXCEPTIONAL CHILDREN

Curriculum-Based Measurement in Middle and High Schools: Critical Thinking Skills in Content Areas

Gerald Tindal and Victor Nolet

Since its inception, curriculum-based measurement (CBM) research and practice have developed in a number of areas, with the primary purpose of validating instructional programs (Deno, 1990). In general, however, CBM has been confined to the basic skill areas and has been studied and implemented in elementary schools. When used within middle and high schools, the emphasis has been on basic skills or general classroom functioning (Espin & Deno, 1993; Tindal & Germann, 1991). In this article we take the major tenets of CBM and extend them into content area learning, the primary concern of middle and high schools. To develop this extension adequately, though, we must consider the essential features of CBM and compare it with other forms of measurement currently represented in the wave of alternative assessments.

CBM BACKGROUND AND PREMISE

Fuchs and Deno (1994) recently highlighted the important distinction between the curriculum used with instruction and the curriculum used for documenting student performance. In that article they argue that, although most educators believe an inherently isomorphic relationship between teaching and testing is needed to ensure content validity, the linkages actually may have to be defined more loosely to ensure formative evaluation with a focus on instructional decision-making. In their argument they proceed with a consideration of domain, generally defined as the materials sampled within curricular/instructional versus tests/measurements. In the end they conclude with the following essential features of CBM.

1. Repeated testing over time is needed on material of comparable difficulty to avoid problems in quantifying changes in performance without confounding instructional changes with assessment changes and to ensure that teachers can ascertain maintenance and generalization of knowledge and skills.
2. Valid outcome indicators allow teachers to answer the question of whether instruction is leading to improvement or has to be adjusted.
3. Qualitative feedback is available to supplement quantitative summaries and help teachers decide not only when to make changes but also to determine what modifications to make.

Gerald Tindal is an associate professor in the Department of Special Education at the University of Oregon. *Victor Nolet* is an assistant professor in the Department of Special Education at the University of Maryland.

This logic and subsequent analysis is critical in understanding any potential applications of CBM to content areas in middle and high schools. To develop an assessment system with these three essential features, the definition of domain must be addressed, as well as sampling plans for focusing both the content and the student responses.

Traditionally, content areas are defined in terms of information, eventually being reduced to topics. Perusal of any major science or social studies text generates the following units, respectively. A seventh grade *Life Science* text (Barr & Leyden, 1989) has topically organized units on the science of living things, diversity of living things, human life, heredity and change, and the environment. Of course, further topical breakdowns are made within each of these units or chapters.

For example, the first unit above includes two chapters on the methods of science and studying living things; the second unit on diversity includes chapters on the life of the cell, classifying living things, plants, and animals. A typical seventh grade social studies book on *Latin America and Canada* (Stoltman, 1989) includes units that are regionally organized around the First Americans, Canada, Mexico, Central America and the West Indies, and South America. These units have more refined breakdowns of topics and subtopics dealing with land-resources-people, past to present, the area today, and facing the future, all of which are addressed in a repetitive format for each unit.

In a corresponding manner, the way we test students has followed a similar organizational strategy, giving rise to an emphasis on criterion-referenced measurement. Over the past three decades "teacher-made" tests and curriculum-based tests have been premised on defining the domains of instruction and then establishing sampling plans for selecting appropriate informational "bits." The last step in this process has been to establish a mastery score or level of acceptable performance. This view of measurement has dominated our schools for 20 years. Only recently, with increasing emphasis on alternative assessments, have some of the assumptions of this testing program been questioned.

To develop CBM in content areas, the domain for sampling content must move from being stimulus-defined (either in terms of the informational "bits" or the format of the response) to one that is response-defined in terms of content knowledge and strategic skill. In all previous work on CBM, the focus has been on student responses first and, subsequently, on consideration of the materials used in generating the response.

For example, in reading (Deno, Mirkin, & Chiang, 1982) the research initially validated a response (among alternatives such as reading words from a list, selecting single words from a passage, producing single words within a passage, reading connected words from a passage). In writing (Deno, Mirkin, & Marston, 1980) the number of words written was eventually validated as the best measure among several alternatives. Finally, in spelling (Deno, Mirkin, Lowry, & Kuehnle, 1980) the number of correct letter sequences produced in spelling words dictated from a list was validated among several other measurement systems.

In all these instances the focus of validation was on the response, with subsequent studies conducted on the impact of stimulus materials and sampling plans for generating tasks. In like manner, therefore, the approach we take is to focus on the response (not just the format but, more importantly, the intellectual operation) as the critical feature for generating a measurement system that has the potential for being repeated over time, addresses instructional effectiveness, and allows both quantitative and qualitative outcomes to be summarized in determining what to teach and when to modify instruction.

OVERVIEW AND PERSPECTIVES OF CRITICAL THINKING IN CONTENT

Before considering operational definitions, some of the critical features of "thinking skills and problem-solving" have to be addressed. In defining "knowledge" we want to

FOCUS ON EXCEPTIONAL CHILDREN

ISSN 0015-511X

FOCUS ON EXCEPTIONAL CHILDREN (USPS 203-360) is published monthly except June, July, and August as a service to teachers, special educators, curriculum specialists, administrators, and those concerned with the special education of exceptional children. This publication is annotated and indexed by the ERIC Clearinghouse on Handicapped and Gifted Children for publication in the monthly *Current Index to Journals in Education* (CIJE) and the quarterly index, *Exceptional Children Education Resources* (ECER). The full text of *Focus on Exceptional Children* is also available in the electronic versions of the *Education Index*. It is also available in microfilm from Xerox University Microfilms, Ann Arbor, MI. Subscription rates: Individual, $30 per year; institutions, $40 per year. Printed in the United States of America. Second class postage is paid at Denver, Colorado.
POSTMASTER: Send address changes to:

Love Publishing Company
Executive and Editorial Office
1777 South Bellaire Street
Denver, Colorado 80222
Telephone (303) 757-2579

emphasize "authentic thinking tasks," to reflect problem-solving and critical-thinking skills that are content- and context-oriented. In the process of solving problems and exhibiting critical thinking, information has to be manipulated. Specific events and content have to be part of the task so we will avoid teaching students "how to think without troubling them to learn anything worth thinking about" (Cheney, 1987, p. 5). In establishing CBM in content areas, we have borrowed from three lines of research.

First, we need to organize information into a sense of breadth and depth. At the broadest and most superficial level is *content knowledge*, which reflects general information that is commonly known and where most individuals are in terms of subject matter expertise. The next level is *domain knowledge*, in which the information base is more developed and the relational network more complex. Most clearly, content-area teachers have domain knowledge. The final level of information breadth and depth is *discipline knowledge*, in which information is considerably richer and attainable only after years of rigorous training and practice. Discipline knowledge is a subset of domain knowledge and typically is highly specialized information that is organized hierarchically.

Second, rather than viewing content as simply information, we consider it within a continuum from declarative to conditional and procedural knowledge (Alexander & Hare, 1989; Paris, Lipson, & Wixson, 1983; Ryle, 1949). *Declarative knowledge* is defining and describing information. *Procedural knowledge* focuses on using information within a problem-solving paradigm. Finally, *conditional knowledge* is reflected in the timing and conditions within which information is presented and used (Alexander, Schallert, & Hare, 1991). Similarly, Skemp (1978) distinguishes between relational understanding (knowing what to do and why) and instrumental understanding (knowing rules without reasons). We believe an appropriate balance needs to be struck between content knowledge (what is learned) and strategic knowledge (how it is manipulated), a distinction highlighted by Alexander and Judy (1988). To understand the relationship between these two types of knowledge in developing cognitive assessment tasks, they summarize four issues:

1. Domain-specific knowledge is required for efficient and effective utilization of strategic knowledge. To develop effective arguments and solve problems, students need specific information represented in their arguments and solutions. As students' content and domain knowledge increases, improvements in strategies also are likely to improve.
2. Inaccurate content knowledge may interact with learning, at times inhibiting or interfering and at other times setting the occasion for focusing on important content information and strategic reasoning. In either case a continuum is likely to exist, further supporting Fuchs and Deno's (1994) argument against mastery measurement.
3. Strategic knowledge facilitates acquisition and utilization of domain-specific knowledge: "Knowing" is probably necessary but not sufficient in solving problems and "thinking critically." Students need to learn how to access information, plan responses, organize information, self-monitor performance, and link information within arguments, all of which should have an impact on their learning content and domain knowledge. The corollary also is important: Ill-informed use of strategies probably impedes learning. Students may need to be taught more explicitly the relationships between content and strategy domains.
4. Competent performance is characterized by students perceiving the relatedness in domain and strategic knowledge across domains and tasks. "As learners acquire more knowledge, they also seem to acquire the ability to abstract or mentally represent a given problem. Furthermore, more knowledgeable individuals categorize or classify problems on the basis of their underlying structures. . . . More capable learners perceive the relatedness of seemingly diverse tasks or domains and use that relatedness to guide their performance" (Alexander & Judy, 1988, p. 394).

Third, we consider taxonomies of student responses that have been introduced over the years since Bloom's seminal work in 1956. Presently, considerable controversy exists over their varying levels of completeness, only two of which are summarized below to illustrate problematic issues.

Bloom, Engelhard, Furst, Hill, and Krathwohl (1956) first proposed a six-category taxonomy for placing cognitive tasks on a continuum. The categories consisted of recall, comprehension, application, analysis, synthesis, and evaluation. The problems noted with this system have been the reliance on unobservable intellectual operations (Williams & Haladyna, 1982), unreliability in classifying items and tasks (Seddon, 1978), and the lack of attention to the type of information being assessed (Tindal & Marston, 1990).

Another system, used by Stiggins, Griswold, and Wikelund (1989) in their investigation of four different types of classroom thinking skills assessment (teacher-developed paper-pencil tests, curriculum-embedded tests, written assignments, and oral questions during instruction), is based on Quellmalz's (1985) taxonomy, with the following categories and definitions:

— recall (remembering key facts, definitions, concepts, rules, and principles)
— analysis (divide a whole into component parts)

— comparison (recognize or explain similarities and differences)
— inference (including both deductive and inductive reasoning)
— evaluation (judging quality, credibility, worth, or practicality).

Again, no reliability information has been reported in terms of classification of assessment tasks, and the type of information or knowledge form is left out of the taxonomy.

In summary, in developing CBM systems in the middle and high school focused on critical thinking in content areas, we have oriented our work around three issues: depth/breadth of knowledge, balance between content and strategy, and the need for a taxonomy to help structure intellectual operations (student responses) within problem-solving tasks. We have framed these three issues around the need for a time-series view that reflects changes in performance.

ESSENTIAL FEATURES OF CRITICAL THINKING ASSESSMENTS

In our conception of critical thinking skills, we include four essential attributes, each of which can be considered as defining attributes (Prater, 1993). They differentiate all examples within our concept from other examples that others often employ. Curriculum-based measurement of critical thinking must consider the knowledge forms and structures within the (content) informational base. A taxonomy of student responses is needed to help define comparability across tasks and over time, with both content and strategy represented in the responses. A domain must be defined for sampling information and integrating curriculum, instruction, and assessment. Finally, a specific response format must be established for actually conducting the assessment.

Knowledge Forms

One of the first essential attributes of critical thinking is the knowledge form upon which the "thinking" operates. We limit this term to four major classes: (a) facts, (b) concepts, (c) principles, and (d) procedures.

Facts are simple associations between names, objects, events, and places that use singular exemplars. The distinguishing feature of a fact is that the association is limited and not generalized across a range of names, objects, events, and places. Facts are difficult to remember without an organizing scheme to link them. Yet they are the basic building blocks to more complex information and are necessary, for example, in developing a vocabulary that can be used to work with concepts and principles.

Concepts are clusters of events, names, dates, objects, and places that share a common set of defining attributes or characteristics. A concept may be thought of as "a category of experiences having a rule which defines the relevant category, a set of positive instances or exemplars with attributes and a name (although this latter element is sometimes missing)" (Martorella, 1972, p. 7). In this definition rules provide the basis for organizing the attributes of the concept. These attributes (some of which are defining and essential and others of which are variable and not essential) in turn provide the criteria for distinguishing examples of the concept from nonexamples (Prater, 1993).

Principles are relationships among different facts or concepts, more often the latter. A principle usually represents an if-then or cause-effect relationship, although this relationship may not be stated explicitly. A principle generally involves multiple applications in which the fundamental relationship among the relevant concepts is constant across virtually all examples of the concepts.

Procedures and strategies consist of a sequence of activities and algorithms used in solving problems, gathering information, and achieving goals (Anderson & Liu, 1980). Our earlier discussion of procedural and conditional knowledge (a la Alexander, Schallert, & Hare, 1991) and strategic thinking applies here. The focus on procedures is "when" and "how" to manipulate information so that problem-solving is both effective and efficient.

Taxonomy of Student Responses

Debate over taxonomies is not likely to translate into practical help for teachers, and any one of the taxonomies may well serve the eventual purpose of capturing student thinking. In the following taxonomy, which we have adapted from Roid and Haladyna (1982), our goal is to achieve the most practical and operational procedures with the least amount of inference. In the definitions below the focus is on a relatively narrow and concrete set of behavioral classes within which students can respond. Student responses refer to the behavior employed in using or manipulating knowledge forms, with five major types. The first two are a combination of three different operations proposed by Roid and Haladyna.

1. *Reiteration-Summarization.* The first component (reiteration) is *verbatim reproduction of material that was taught previously.* With reiteration, emphasis is on *verbatim.* The wording in the student's response must be nearly identical to that presented in the curriculum and within instruction. With the second component (summarization) we consider *generation or identification of a paraphrase, rewording or condensation of content presented during*

instruction. Emphasis here is on previous presentation of material. Summarization involves remembering information rather than manipulating it. In the end we have grouped these two together because of inherent difficulties in distinguishing them in the linkages across curriculum—instruction—assessment.

2. *Illustration.* In this intellectual operation we focus on *generation or identification of a previously unused example of a concept or principle.* Emphasis is on *use* of an example that was not presented in the curriculum and within instruction. In this respect the student is expected to employ information about the attributes of a given concept or principle rather than to simply remember whether an event does or does not exemplify a knowledge form.
3. *Prediction* involves *description or selection of a likely outcome, given a set of antecedent circumstances or conditions that has not been encountered previously.* Again, emphasis is on the *use* of information in a novel context rather than remembering a response from previous instruction.
4. *Explanation* is the *description of the antecedent circumstances or conditions that would be necessary to bring about a given outcome.* Explanation is the reverse of prediction. The student must use information about a concept or principle to work backward from the circumstances presented and tell what happened to create it.
5. *Evaluation* is *careful analysis of a problem to identify and use appropriate criteria to make a decision in situations that require a judgment.* Evaluation focuses on decision-making: The student first must recognize or generate the options available and then use a set of criteria to choose among them.

These five learning tasks interact with the knowledge forms described above so that some operations may be more realistic than others with some knowledge forms. Reiteration-summarization (which we have grouped together, given the difficulties of knowing whether the response is exact or a paraphrasing) can occur with all three knowledge forms. Students engage in these forms of behavior when asked to recite facts, recall definitions of concepts, restate lawful relationships, or follow procedures.

The remaining four can be applied only to concepts, principles, and procedures. Illustration requires an individual to recognize or generate previously unused examples of a piece of information. (A fact cannot be illustrated because it consists of a single simple relationship between constructs.) Prediction of a concept occurs when a student is given some but not all of the defining attributes, and then uses them to infer other attributes of the concept or inductions-deductions of a principle. Prediction of a principle occurs when a student is given a previously unused description of a situation that has the antecedent conditions (causes) of a relationship embedded in it and is able to use that information to identify the most likely consequences (effects). Explanation, the opposite of prediction, is used when a student is given an outcome (effects) and some initial state and then determines the conditions (causes) required to achieve that outcome. Evaluation tasks require two basic steps: (a) select criteria, and (b) make a judgment based on these criteria. The judgment has to be supported by the criteria. The emphasis of problem-solving, however, may be in explicating the criteria, making decisions, or both together. Implied in making decisions and supporting them is to argue not only for a choice but to argue against the nonchoice.

Irrespective of the specific intellectual operation, the last four represent manipulations of information, which separate this work from more traditional views of comprehension in social studies and science. For example, considerable research and writing has appeared around text structure (Armbruster & Anderson, 1984; Armbruster, Anderson, & Ostertag, 1989; Meyer, Brandt, & Bluth, 1980). In this line of investigation, writing (of textbooks) and comprehension (by students) has been studied so that students can become more proficient in understanding information. The question, however, is whether any comprehension tasks based on these analyses are any different than summarizing-retelling. If no demands are made in the assessment task to manipulate the information or modify and rearrange it, can we legitimately ascribe the outcome to critical thinking (Nickerson, 1985)?

> In short, we want our students to be able to answer questions that were not part of the lesson. . . . Understanding [therefore] refers to a student's ability to creatively use presented information to solve transfer problems. (Mayer, 1989, p. 43)

As prompts are structured, we have made a clear distinction between only five intellectual operations (reiteration-summarization, illustration, prediction, explanation, and evaluation). We believe the latter four primarily reflect "critical thinking." In our view of critical thinking, requiring students to illustrate, predict, explain, or evaluate reflects manipulation of information.

Classroom Contexts and Domain Sampling

Another essential feature of critical thinking skills, domain, is "a specific set of skills or body of knowledge associated with an instructional intervention. This domain also may include a continuum of competence in using the skills or knowledge" (Tindal, Nolet, & Blake, 1993, p. 13). A domain defines the potential pool for sampling information

and response tasks. The sampling is obviously contingent, however, upon adequate definition and focus on knowledge forms and an appropriate match with the intellectual operations. Although these two attributes are necessary, they are not sufficient, for in the end, a judgment of a student's performance must be made on multiple data sources, demanding some type of sampling plan that converges across curriculum, instruction, and assessment.

Our definition of domain has two components. *First*, a domain should be linked across three arenas: curriculum materials, specific instructional interventions (both tactics and strategies), and assessment formats. For example, a domain cannot simply pertain to "Chapter 13 in the social studies book," or "three months of instruction on writing compare-contrast essays." These examples involve only curriculum or instructional foci, respectively; consideration also must extend to the assessment dimension.

Second, our definition of domains involves the *use* of skills or information as well as the *information* itself. The taxonomy of intellectual operations presents a range of responses that students could make in using information. In the end, a time-series of tasks is presented to ascertain changes in the use of concepts and responses the students made, with both quantitative and qualitative summaries and interpretations made from their responses (Fuchs & Deno, 1994).

Although considerable attention has been devoted to analyzing either the curriculum or instruction as a means for determining content validity, we approach the issue from the perspective of Frederiksen and Collins (1989), who use the term *systemic validity* to depict the relationship between curricular and instructional changes that arise from using specific measures. In essence, the three contexts of curriculum, instruction, and assessment are considered in defining a domain and subsequent sampling plan. When all three of these attributes are considered together, we have a three-dimensional cube, with any one cell defined by the intersection of the three planes. In our emphasis on critical thinking, we highlight tasks in the shaded boxes of the cube depicted in Figure 1.

Response Format

The final feature of a critical thinking skills assessment is the response format (Figure 2), which, though inherent in any assessment task, helps us define some nonessential attributes. Currently, considerable debate exists over the whole range of alternative assessments. Many new terms are being used to focus on "performance" tasks and "authentic" conditions. The essential logic is that unless students produce the response in a problem space that mimics the "real" world, the assessment is suspect. We take a far

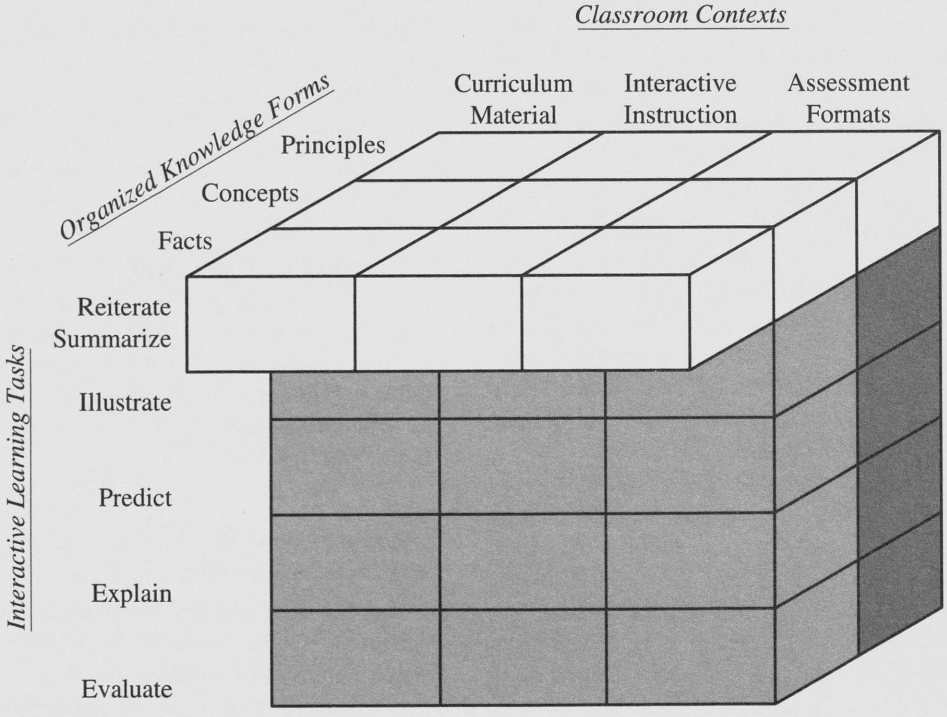
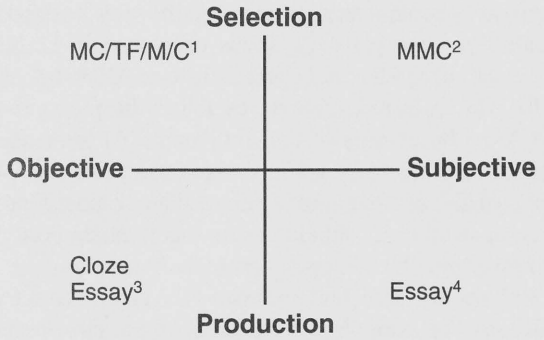


FIGURE 1
Three Features of Critical Thinking Skills Assessment



1. MC = multiple choice/TF = true-false/M = matching/ C = classification
2. MMC = multiple-multiple choice
3. Cloze = one word completion/short answer essay
4. Essay = long answer essay

FIGURE 2
Dimensions for Formatting Student Responses

more moderate and empirical point of view. In the next section we address a host of issues and present a variety of outcomes reflecting a broader and more encompassing perspective. To ensure a common language, however, some definitions may be needed:

Selection Versus Production

Traditional measurement texts have divided student responses into two classes in which students either select the response (usually including multiple-choice, true-false, matching, or classification) or produce it (usually composed of one word, short answer, or essay/extended answer). We assume that critical thinking tasks can include either selection or production; however, because of our emphasis on CBM, we focus on production tasks.

Objective Versus Subjective

Often an inference is made that implies selection tasks are objective and production tasks are subjective. In our scheme these two dimensions are independent and any combination of both dimensions can be represented in any assessment task.

Although most of the quadrants are probably familiar, the selection-subjective quadrant has appeared only recently. Multiple-multiple choice (MMC) tasks were introduced in the Illinois large-scale assessments a few years ago in an effort to get students away from considering all comprehension answers as absolutely correct or incorrect (Valencia, Pearson, & Chapman, undated). Basically, different tasks were used with multiple-choice items in which students rank-ordered the most to the least logical outcomes or selected the three most likely outcomes. Because the selections could vary in terms of plausibility, a subjective scoring system is warranted.

The other three quadrants are self-explaining. Traditional multiple-choice tests are examples of objective-selection; only one of the four distracters is usually correct. Cloze techniques, in which a word is missing from a sentence and the student is required to fill in the (semantically/syntactically) correct word (see McKenna & Robinson, 1980, for an annotated bibliography) and short-answer essays reflect objective-production tasks. Finally, long essay responses are usually scored subjectively, but as we propose, a number of objective counts also can be used to summarize performance. In our CBM focus we consider both quantitative and qualitative dimensions of production tasks.

Direct Versus Indirect Measurement

These terms are often confused with production and selection, respectively. Production tasks are assumed to be direct measures, and selection tasks indirect. Yet, all measures of "critical thinking" may be indirect. The only direct measures (WYSIWYG, a term devised by Apple® computer to mean What You See Is What You Get) may revolve around basic skills or quantifiable indices (such as syntactic features in written responses and countable indices of concept usage). We ignore this dimension in developing CBM tasks, simply assuming that our measures are direct samples of writing and speaking and indirect measures of "critical thinking."

Reliability Versus Validity

This issue is tangled with the new alternative assessments arising in the form of portfolios and other performance tasks. In general psychometric terms, all measures are required to provide reliable (consistent or stable) estimates of performance before they can be considered valid (truthful) indicators.

Ironically, the reliability of performance estimates in many alternative assessment systems has been low. For example, the Vermont portfolio assessment program reported low consistency in the interjudge scoring of the writing and math samples; coefficients ranged from .33 to .43 (Koretz, Stecher, Klein, & McCaffrey, 1994). The Oregon Statewide Assessments revealed few exact agreements on six writing traits (ideas-content, organization, voice, word choice, sentence fluency, and conventions). In general, less than 50% of the writing samples were given the same exact score by more than one reader (and this was true across all the grades 3, 5, 8, and 11) (Oregon Department of Education Technical Report, 1989–1992).

Increasingly, validity is being redefined in terms of decision-making—the value and social implications that result from measurement usage. In Messick's (1989) view, validity is a uniform term in which test and measurement informa-

tion is interpreted in the context of evidence and results, giving rise to a variety of consequences. That is, not only must tests have supportive evidence for proper interpretation, but they also must have relevance and utility for the purposes to which this information is being applied. Furthermore, test interpretation is provided value only via the implications of those interpretations and the social consequences that inherently arise with use of the test. Clearly, performance assessments demand a richer and broader view of validity for appropriate test interpretation and use (Moss, 1992).

Although the general argument is that alternative assessments are more meaningfully anchored to "real-world" contexts, and therefore are more valid, little agreement can be reached in making judgments of student quality. In developing CBM systems in content areas, we have addressed directly the issue of reliability and validity, as was done with the original basic skills research conducted in the early 1980s (see Nolet & Tindal, 1993, and Tindal & Nolet, in press, for some of the initial validation research).

In summary, critical thinking focuses on information breadth and depth that has an element of strategic reasoning. Specifically, four attributes reflect critical thinking:

1. A range of knowledge forms that help frame the informational base.
2. Any of several student responses or intellectual operations that focus on manipulating the information.
3. A domain that integrates curricular texts, instructional events, and assessment tasks.
4. A response format that varies across two dimensions of objectivity-subjectivity and selection-production.

In the next section we present some guidelines for operationalizing all these features of critical thinking skills into actual prompts to which students respond. We primarily consider long- and short-answer written (essay) responses, but all of the issues we discuss apply to assessments that require other responses, such as interviews and graphic displays (for example, maps or diagrams). Nevertheless, we always proceed from an analysis of the knowledge forms (facts, concepts, principles, and procedures) and intellectual operations (reiterate-summarize, illustrate, predict, explain, and evaluate) within three contexts (curriculum, instruction, and assessment). In the last section, we present a year-long time-series view of a student in an eighth grade science class, providing both quantitative and qualitative summaries of performance.

DEVELOPING CBM PROBLEM-SOLVING PROMPTS

Like blueprints for building construction, assessment prompts should have a structure that identifies specific informational content and response. At the very least, prompts should have two parts: (a) focus on a specific knowledge form, and (b) intellectual operation stem. A context often is added via an introduction to establish the problem space within the broad area of content (topic). At least one sentence should introduce a knowledge form to be used within this specific context and content. The response demand made upon students should ensure that they are performing the correct intellectual operation in their answer.

With some intellectual operations the architecture may be elaborate. For example, with evaluation prompts a sentence may introduce the knowledge forms and may present at least two sides of an issue: advantages and disadvantages, similarities and dissimilarities, positive and negative influences, and so forth. The purpose of the evaluation stem is to prompt the student to consider both sides and choose between these alternatives. Or, with a prediction item, a middle sentence may have to include a time sequence to convey the correct trajectory of events so the eventual prediction stem can depart from a well-established point.

Irrespective of elegance or complexity, eight considerations lie behind the prompt, which may be considered in any of the three arenas of the domain (curriculum, instruction, and assessment).

1. Pivot words are used to direct student responding.
2. Choices are presented with equal valence.
3. Administration formats are instructionally relevant elements of the domain.
4. Student response may be scaled either quantitatively or qualitatively.
5. Elaboration is explicitly solicited.
6. Response scales are sensitive.
7. Response strategies are implied.
8. Scores are embedded in the prompt.

All of these issues may be part of the assessment; they also are likely to be part of curricular-instructional events in establishing systemic validity.

Pivotal Words in the Stem

Specific pivotal words are the most important influences in the prompt. They have to be chosen with care because they are intended to lead the student to a specific intellectual operation. Generally, these words are verbs, though they may be adjectives (with illustration) and verb-objects. Following are examples of pivotal words for each type of intellectual operation, with the key phrases underlined:

1. Reiteration: *Repeat the exact words, recite, state the definition, give a verbatim* description. Example: "Recite the preamble to the Bill of Rights."

2. Summarization: *Paraphrase* the content, *retell* what you read, *summarize* what was said, *describe* the main issue. Example: "Describe the main problem engineers encountered when they first tried to extract oil from shale."
3. Illustration: Provide an *example*, present a *comparable issue*, relate an *analogous incident*. Example: "From your own experience provide an example of the consequences of procrastinating."
4. Evaluation: *Decide* which *alternative*, determine the *correct choice*, *consider* which *option*, *compare* and *determine*, *select* one, and *justify*. Example: "Of the alternatives to the use of fossil fuels listed above, select the one you think would work the best, and justify your answer."
5. Prediction: Tell *what will happen*, *make a prediction*, *guess an outcome*, describe *subsequent events*. Example: "If all the plankton die in a lake, describe what will happen to the other organisms living in the lake."
6. Application: *Explain the outcomes*, *give some reasons*. Example: "Use what you have learned about plate tectonics to explain the existence of the Cascade mountain range."

Equality of Choices

A potential source of error, and thus a threat to reliability and validity, is the amount of "bias" contained in a prompt. Students should not be tipped off by the wording that one choice or type of response is better than another. They should be able to demonstrate what they have learned by making effective choices and arguments on their own. The prompt must be worded so all choices embedded implicitly within the text are equal. Although this requirement sounds obvious, a number of subtle influences may make one choice more relevant or "answerable" than any other. Following are two examples:

In describing the content (knowledge forms) and context, more information may be provided for one specific concept than another. For example, if an essay about two forms of economy (*command* and *market*) has a disproportionate amount (and specificity) of information about economics, students' answers may be influenced. The breadth of the concepts may be different so one choice has fewer opportunities for an extended answer. An essay may cover two biomes (*rain forest* and *tundra*) and the natural resources converted to human usage within each. Clearly, fewer natural resources are exploited from the tundra, thereby prompting students to answer more often with reference to rain forests.

Instructional Relevance

Like all end-of-unit assessments, the assumption often (and incorrectly) is that the information covered on the test has been addressed in the course of the curriculum and instruction. Little research, however, has been completed on this topic from an instructional viewpoint. As stated earlier, the research that has been done has focused primarily on the overlap of the curriculum and published achievement tests.

The issue, then, is how closely the textbook is aligned with instruction delivered in the classroom. And, with the framework we have provided (using a three-dimensional analysis of knowledge forms by intellectual operations by classroom contexts), this issue becomes much more complex than just counting the number of times a vocabulary word or an informational item is included in instruction and on an assessment. Rather, the specific knowledge form and its use as an intellectual operation must be clear across these dimensions of curriculum, instruction, and assessment.

For example, a concept such as fossil fuels may be emphasized heavily within instruction, using "prediction" (of positive and negative consequences from heavy dependence upon fossil fuels). If the end-of-unit measure, however, contains items and tasks requiring students to simply summarize their uses, the focus on critical thinking may be negligible and the assessment results minimal. Most of the problems posed within the assessments are transfer tasks and may not have been taught or presented directly.

Administration Format

When administering assessments, a key issue involves the *level of prompting* and the *task demands*. *Level of prompting*, or the amount and kind of supplemental help students receive to organize their response, can be great or minimal. For example, the teacher may read the extended text for a prompt to students, highlighting key words or phrases along with verbal prompts to recall the content of instruction (e.g., "you need to think about the part of the chapter that covered the nitrogen cycle . . . remember how a certain chemical catalyst got the cycle started . . ."). This kind of administration would be prompted heavily. In contrast, the teacher may distribute a simple prompt and direct students to read it and begin when they are ready to write. These procedures would reflect almost no prompting.

Task demands refer to the mode of response required of students. Because written essays leave a permanent product that can be scored and evaluated later, they are easy to use. They also require students to be (somewhat) facile in writing. If the purpose of the measure is to understand students' performance using a specific intellectual operation, their response should not be unduly influenced by the manner in which they respond. For example, a student may have exhibited much more (or less) knowledge simply because of the task demands (e.g., writing out the response instead of talking it out). The outcome, then, must be considered as a mix of the intellectual operation and the task demands; it is impossible to sort them out separately.

Scale of the Response

Student responses may be scaled either quantitatively or qualitatively. The extent to which students know how they will be judged may influence how they perform. Actually, arguments can be made both for telling them in advance and for withholding that information from them. The choice depends upon the purposes of the assessment.

For maximal instructional application, students should be told in advance or as part of the prompt how their performance will be evaluated. For example, they may be directed to "use as many of the important concepts" or "link the important concepts into a major principle" because that is what is valued. Students even may be given the actual qualitative scale to use in constructing their responses (so they can see the anchors for a low response versus a high response).

On the other hand, for maximal generalizability to a larger context or group (e.g., across time, teachers, classes, or grades), students may not be told anything about the scoring systems; rather, they simply may be directed to do their best. In these instances the assessment task should reflect the fact that in generalized settings students will encounter a variety of prompts and response demands. They would need to be able to extract the key elements of the task and respond appropriately.

Explicit Reference to Elaboration

Generally, students receive relatively little practice on writing or speaking tasks. Therefore, an extended essay using a problem-solving situation may result in a narrow distribution of scores (centered near the bottom, with only a few high performers). Students probably need to be explicitly directed to "justify the answer," "give as many reasons as possible," "provide as much supporting detail from the chapter as they can," and so forth. These extra prompts should communicate to students that they have to do more than write a minimal response. If extended responses are used regularly in a classroom and students are taught the criteria by which their responses are judged, this issue is likely to diminish in importance over time.

Response Strategies Within Student Performance

In every facet of classroom interactions, students need strategies to perform successfully. For example, when teachers are lecturing, these strategies may involve note-taking: During independent reading, students may need to rely on any of the various SQ3R-type strategies to help them interact meaningfully with the text. A host of other test-taking strategies has emerged over the years to help students perform better. This last issue is what has a bearing in the use of extended production tasks, particularly in creating a prompt that generates a response capable of being evaluated reliably and validly.

Student responses to essay tasks may be influenced by a number of components such as knowledge of the content, background knowledge, and writing process skills. To ensure that the task can generate reliable estimates of critical thinking performance, other influences have to be controlled. Following are some suggestions that should help focus the task on the knowledge forms and the intellectual operations of interest.

1. In more elaborate prompts, try to include enough broad and commonly known background information so that all students have an equal chance of responding to the content of the domain only.
2. End the prompt with directions on how the response should look, and even include some qualifying criteria. For example, tell students explicitly that, although they should write legibly, their answers will *not* be scored for penmanship or spelling; if they don't know how to spell the word, do their best. Or, to prevent students from including most of the information from the prompt, direct students to "focus on the content of instruction and use reasons from their knowledge of their subject, not just the information from the prompt."
3. Encourage students to take a few minutes before they begin writing to review the prompt and plan their response. Use a planning sheet to let them organize (and even outline) their ideas before they begin writing.
4. Direct students to work the entire time allotted. Tell them explicitly that, even though they may finish early, they should read their response to make sure they have answered the question adequately.
5. Give a running time notice to help students pace themselves so they do not spend too much time on any single facet of the response. For example, if an evaluation problem requires students to compare and contrast two events (a summarization task), direct them to move on to the part of the task requiring them to make a decision and support it with specific criteria.
6. Provide a motivating statement that allows everyone to perform with equal diligence. Although motivation is a key aspect of student performance for all students, quantifying and controlling it is difficult. The most typical strategy is to tie performance on the task to grades in the course, which may be exactly the wrong way to proceed. For example, if students are told that their answer will be used to form half their grade, some students (e.g., those for whom grades are not motivating, those with such bad grades entering into the task they are flunking regardless of how well they do on this task, and those who have performance contingencies on their grades at home) may respond differently, not

as a function of their knowledge but simply because of the differential effects of the contingencies.

In summary, the prompt should be directed to the specific knowledge forms and intellectual operations of the content and eliminate, as much as possible, influences from background knowledge, administration context that sets the occasion for differential responding, and response process skills (i.e., writing skills).

Embedded Scores Within the Prompt

Although ascertaining how students will respond in advance of the task is difficult, attempting the range of responses anticipated by the evaluator may be critical. That is, a good way to make sure the prompt is working is to respond to it directly, attempting to write an answer at all levels of quality that might appear within the students' responses. By writing out an "answer key," the evaluator may be able to identify glitches and points of confusion that are part of the prompt. This answer key should establish clear differences in a qualitative scale that other (experts) in the field would agree represent differing degrees of performance. And establishing the extreme scores (the least acceptable and the most elaborate answer) is important.

At some point the evaluator may have to distinguish between a legitimate score of zero and missing data. For example, if a student responds to the prompt but is incorrect and illogical, a score of zero may be awarded. With a response that is on a different topic or that clearly represents a misunderstood prompt, however, to consider this response as missing would be more accurate (the task was never really attempted). In the end, this trial scoring key may be used either to frame anchors on a qualitative scale or to provide clues to the evaluator on what to identify in student papers when constructing/selecting exemplar papers (range finder papers).

Summary of Prompt Design Strategies

The major purpose of extended production response assessments, including student essays, is to allow as much flexibility as possible in constructing an answer so both student process and product can be evaluated. Process refers to the manner in which students reflect upon and construct meaning using specific knowledge forms and intellectual operations. It refers to looking at their "thinking" diagnostically and is best considered from a criterion-referenced view. That is, what misconceptions appear to be present? How are different concepts elaborated? What intellectual operations reflect manipulation of the knowledge forms? In contrast, product refers to the outcome, however it is attained. This view is likely to be either norm- or individual-referenced, with the respective purpose being to show relative position or change over time.

Regardless of the focus on process or product, the response has to be interpretable relative to instruction and not unduly influenced by extraneous factors. The list of strategies above should help create worthwhile tasks that are truly classroom-based. If all these suggestions are followed, the inferences made from the data are likely to be more reliable and valid.

EXAMPLES OF SCORING SYSTEMS FOR QUANTIFYING/QUALIFYING CRITICAL THINKING

In applying CBM to content areas, we emphasize the time-series nature of performance: Our primary objective to is to determine whether student performance is improving. If it is not improving, we need to make a change. If it is improving, we need to be certain that (a) our instruction is responsible for the improvements, and (b) even further improvements may not be possible. In either case, then, we should be certain that successive measures are comparable, using the logic outlined by Fuchs and Deno (1994). Another major goal of our measurement system is to be sensitive to low-performing students and changes over time, thereby avoiding some of the problems encountered by Baker et al. (1991), in which all students performed extremely poorly on different problem tasks in history. In actually scoring performance, we have relied upon both quantitative and qualitative summaries.

In this last section, we present a case study as an example. In general, our quantitative count focuses on the knowledge form, and the qualitative judgment focuses on the intellectual operation. As noted in the flowcharts for assigning a quality score of 1 to 5, however, we have reduced the judgmental aspect of this measurement system to a minimum.

Quantitative Scoring

We have taken a relatively simple approach in looking at student learning in content areas by focusing on the use of concepts. If one of our primary goals with teachers is to focus instruction on a relatively few number of concepts and principles and to be more clear and explicit in their usage in the classroom, it inherently follows that we should ask the same of students. By counting usage of concepts and principles, however, we quickly found a number of issues that must be resolved.

The first concern is what to call a concept. Often students use synonyms, ill-defined referents, and even attributes to refer to various concepts. In quantifying concept usage, therefore, a rule has to be established regarding whether to allow all such variants to be counted. Furthermore, any one (essay) response may include concepts from the current or earlier chapters. In the student samples in the last section, we take the most conservative approach and count only ex-

plicit instances of concepts; however, we count all concepts in any given response, regardless of the unit or chapter in which it was presented initially.

The second concern involves duplicated counts of concepts and principles. A short- or long-answer essay response is likely to address several facets of any given knowledge form, none of which is discrete and exclusive of the other. Again, to maximize reliability, we have parsed our concept usage into unique events. Every instance of a concept or principle is counted only if it is within a new context (of other concepts or principles) or presents additional attributes and examples.

Finally, we have addressed the issue of responses being written primarily by counting the number of words and thought-units in the response. In our earlier work we found that density of concept usage may be an important dimension and, therefore, strategic answers have to be considered.

Qualitative Scoring

We have developed flowcharts to step the evaluator through a series of decisions in qualifying the students' responses. The general emphasis of the charts is on increasingly differentiated responses that are both connected (logical) and correct (anchored to the content). We include only prediction, explanation, and evaluation at this time.

Prediction

In this intellectual operation we focus the lowest level of point awards to responses that comprise predictions but not in response to the prompt. Often students will react to a problem and answer a different question logically, showing some signs of a reasoned solution. In moving to higher levels of response, we then focus on logical derivations of the basic premise (or prediction), expanding an essentially binary decision (yes or no) to one of considering the number of elements in the derivation. Finally, in attempting to distinguish a high from a very high response, we focus on the entire argument and its connectedness. Figure 3 is an example of a flowchart to help scorers evaluate student responses.

Explanation

The major feature of any student explanation of why certain outcomes occurred is the decision. Therefore, we begin by anchoring the response to the prompt. Again, students may provide a well-articulated explanation to an outcome different from that presented in the prompt. Moving up the scale of quality, we then ask whether the reasons ("events") are simply present (again, a binary decision of yes or no) or plentiful (multiple). In ascertaining the quality of these reasons, we consider accuracy of the information (unlike predictions, which can be only inferred), and we can go back over the content subject matter to judge accuracy. This last dimension differentiates the high from the very high responses. The flowchart is presented in Figure 4.

Evaluation

The focus on this intellectual operation is the decision and the arguments made both in favor of the choice and against the non-choice. We also quickly link the rationale to reasoned argumentation that essentially becomes content-bound (that is, the information presented in defense of the decision must be correctly related information from the content subject matter. When opinions (only) are presented, the quality of the argument is downgraded. In defending the choice, we move primarily to a reasoned (content-based) argument for one choice and against the other, reflecting our emphasis on argumentation. Following this minimal distinction, we address the multiplicity of reasons as distinguishing high from very high responses. Figure 5 presents a flowchart for awarding points in evaluation responses.

Examples with an Eighth-Grade Student with Learning Disabilities in a Science Class

Actual samples of problem solutions generated by an eighth-grade student with learning disabilities are discussed here. This student received instruction in science from the general education content teacher and supplemental instruction from a special education teacher operating a pull-away program. A collaborative model was used to supplement instruction and tie it to concepts and problem-solving (see Nolet & Tindal, 1994). For each unit a content planning sheet was used to direct both general and special education teachers to a limited set of concepts. Intellectual operations were devised and practiced repeatedly in both classrooms, although the bulk came from the special education teacher. The emphasis of the special education support was on learning key concepts through a series of units, practicing use of these concepts within problem-solving tasks and then summarizing performance both quantitatively and qualitatively.

In the samples in Figure 6, atmosphere had two samples that were evaluated, the water cycle (moisture in the air) had four samples, weather had five samples, an ocean unit had three samples, plate tectonics had seven samples, earthquakes had three responses, and astronomy had three responses. Figure 6 shows the students' responses on problem-solving tasks related to these units throughout the year. Table 1 gives the key for aligning successive tasks by intellectual operations.

Qualitative evaluations were completed by the special education teacher using the flowcharts in Figures 3–5 and the information from Figure 6. Reliability of judgments was conducted by a doctoral candidate. In making the judgments, agreement on the flowcharts was considered in calculating

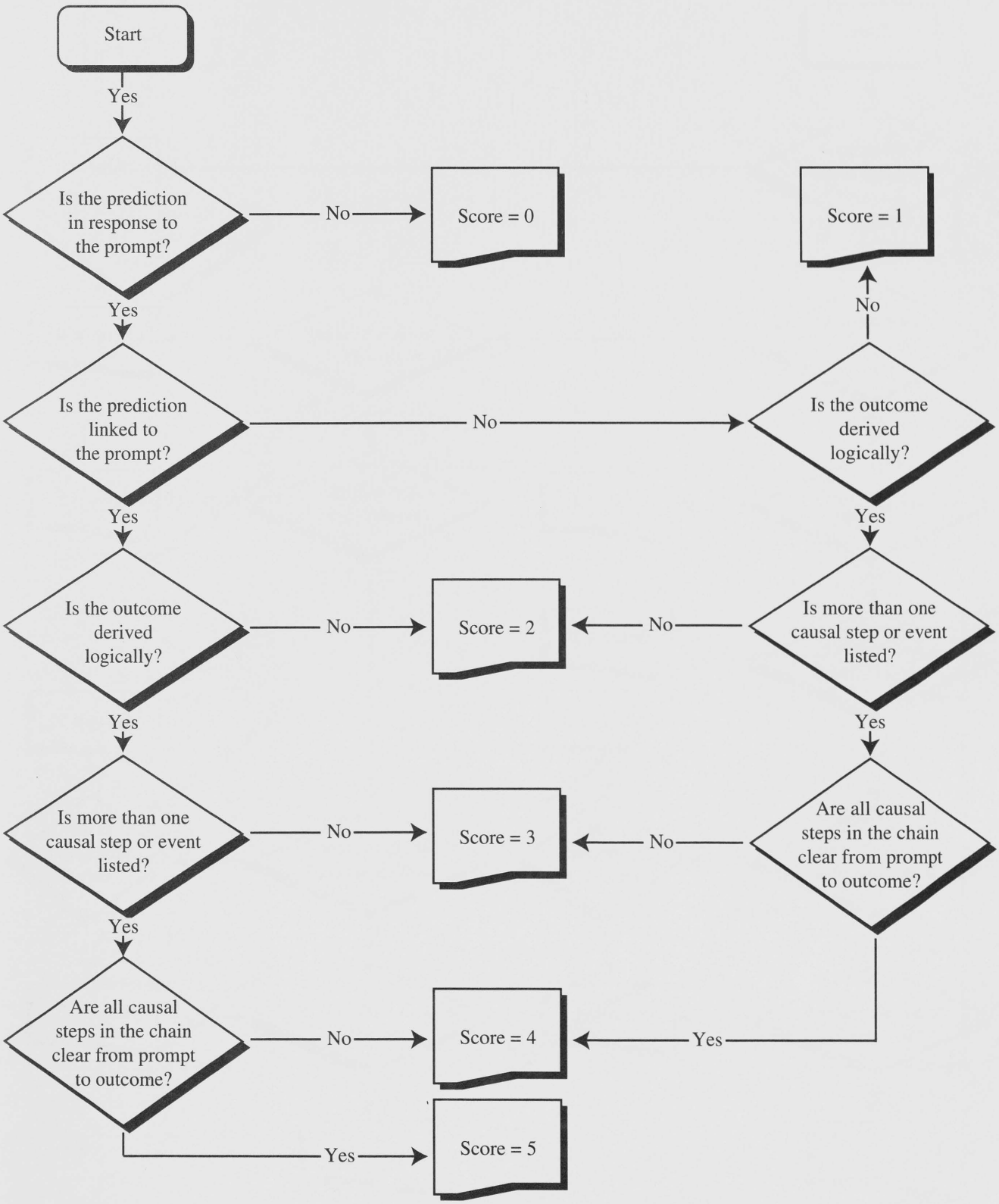


FIGURE 3
Flowchart to Help Evaluators Rate Prediction

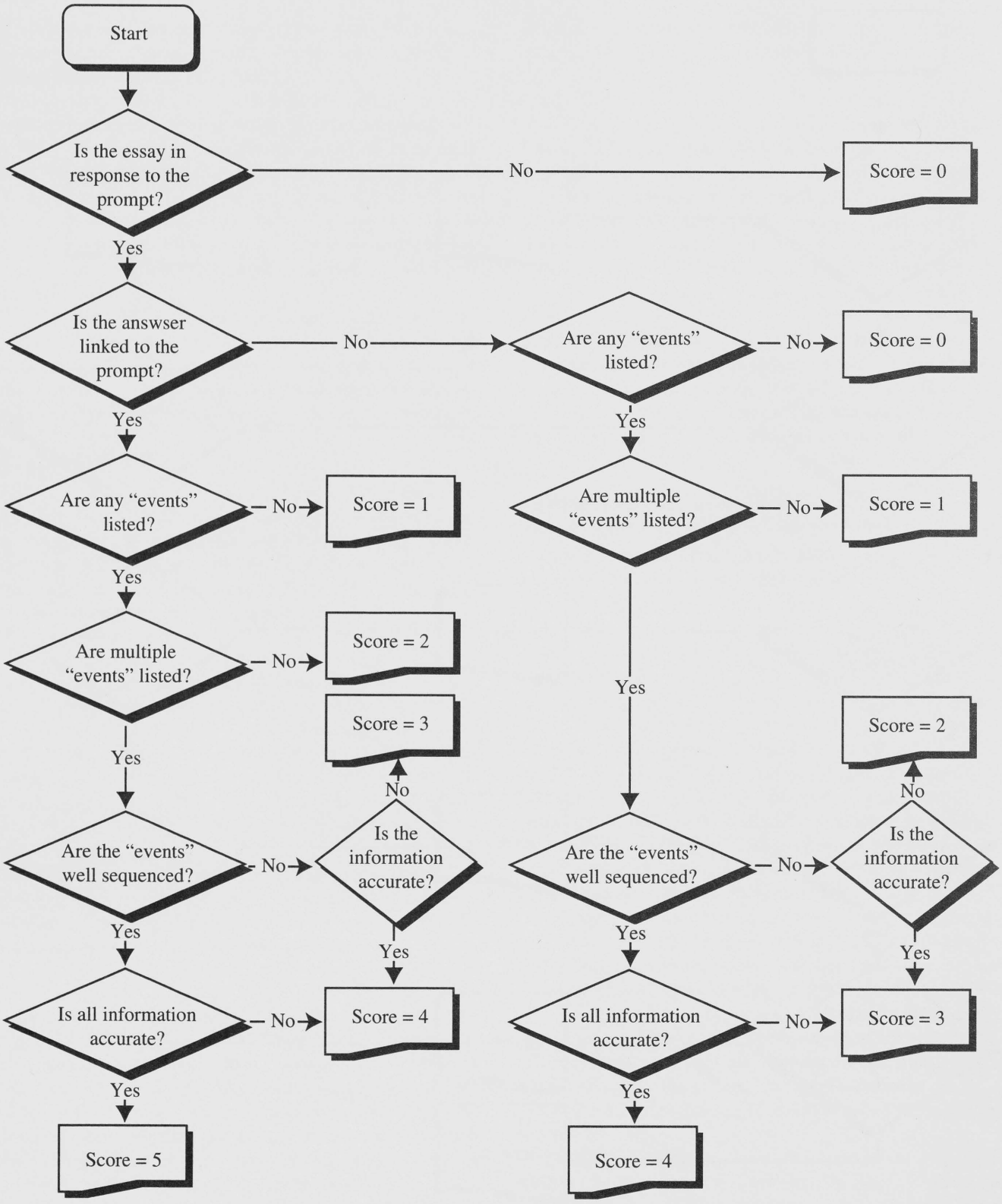


FIGURE 4
Flowchart to Help Evaluators Rate Explanation

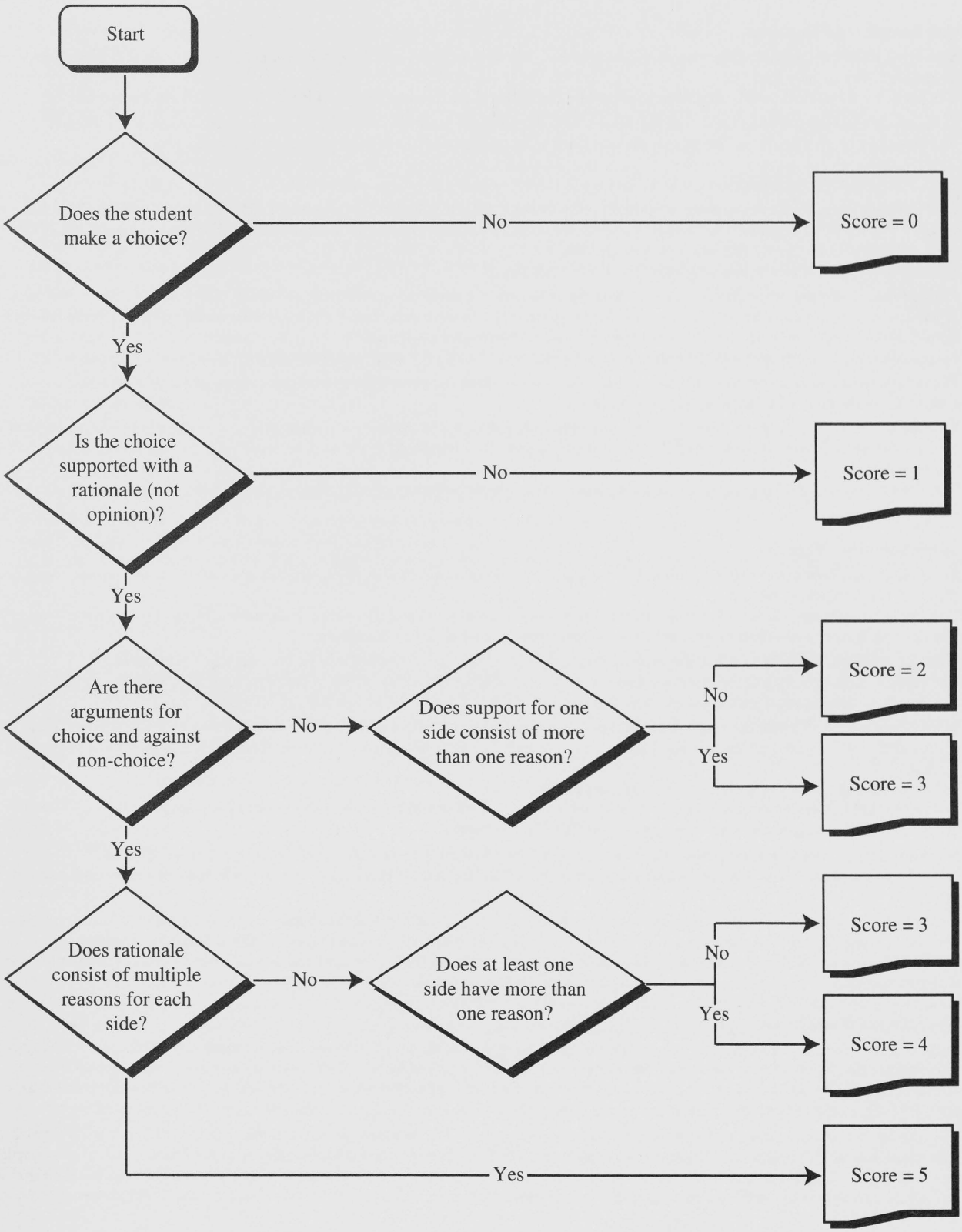


FIGURE 5
Flowchart to Help Evaluators Rate Evaluation

October Samples (Atmosphere)

1. *Suppose a planet in another solar system has a higher gravity than on earth and is colder than earth. Do you think it would have an atmosphere? Why or why not?*
   Yes becomes it has't To have a atmasphere becomes a atmapher bloks to much heat so If it is cold it must have a atmapher.
2. *Is the atmosphere different on earth and the moon? Why do you think so or why not? What affects the atmosphere of a planet?*
   Yes you can't breth The Air on the moon becomes there is no oxegen and There is no gravity so The gases Flot away.

November Samples (Moisture in the Air)

3. *One morning the ground around your neighborhood is wet and saturated with water. The night before it rained heavily. That same rain used to be in the form of groundwater at an earlier time. What do you think happened to change this water from groundwater to rain and back to groundwater again? Explain why you think so.*
   It raind Then The sun evaperated the water it got cot in the cloud.
4. *Pretend that when you leave for school one morning, you notice a small puddle in your driveway. The entire day is steamy and hot. When you get home that afternoon, the mud puddle is gone. Explain what you think happened to the puddle of water while you were at school. No the dog didn't drink it. Use what you know about how water changes form.*
   evaporashon the sun's energe and radeashon heat the water and That is how evaporeshon werks.
5. *The air contains a certain amount of water vapor. What do you think happens to the water vapor as the air cools?*
   It starts to do the sikel of condensation gas to a liquid.
6. *Imagine you have filled a glass with ice cubes and water and left it on the counter for a time. When you return, you notice drops of water have formed all over the outside of the glass. What do you think happened to cause the water drops to form on the glass? Tell why you think so.*
   The glass is cold it is making the air around the glass so water vapor is terning into liquid so it need to hold on to something soit hold on The The glass.

December Samples (Weather)

7. *An air mass that forms over northern Canada is cold and dry. What would be the characteristics of an air mass that forms over tropical Hawaii? Tell why you think so?*
   The characteristics are The cold Front wich came From The Tropical wind and The ocean That made a Front. The dry Front came From The dry winds over The ocean From a nother Iland and From The sands From the hawaii.
8. *What do you think happens to cause the warm air mass to rise or be pushed upward when a cold air mass approaches?*
   The cold air mass is hevy and the warm air mass is lite so The cold air mass stays down because its malicules are slow so they get When they hit the warm Air mass gets bouncte up in The Air because it is lite so That is how a Front happends.
9. *What would the earth's climate zones (tropical, temperate, polar) be like if the earth was not tilted on its axis? Tell why you think so?*
   If the earth was strat up and down it would not change thrugh The hole hear then it Probably would get cold From not geting sun and warmth.
   Poler: would change by not geting heat it would get colder.
   Tempret: would chang by geting maore sun it is more out and more shoning To The san so it would get warmer.
   Tropicle: would chang begeting hoter by geting hit with The sun more.
10. *When a weather front forms, precipitation usually occurs. What happens to cause this resulting precipitation? Explain.*
    The cold air mass moves tuword the Hot air mass the cold hits the Hot and the Hot air gos up and cools and turns into condonsashon and it rains.
11. *A continental polar (cold and dry) air mass is moving slowly across the land. It eventually moves over ocean water. This ocean water is warmer than the frozen land the air mass formed over. What do you think will happen to the air mass? Tell why you think so.*
    It is going to change To wet because it is over water the water is warmer Than The land That it was over so It takes on humidity and it will get warmer.

January Samples (The Ocean)

12. *Suppose a strange bacteria is killing much of the plankton (drifters) in the Pacific Ocean. Many varieties of plankton are disappearing rapidly from the ocean. How do you think other ocean life might be affected by this? Tell why you think so.*
    affects ocean life by the Food web. The Bacteria gets eaten From the zooplanton get eaten by drifters get eaten by nekton and nekton get eaten by bigger nektons like wales. so if bacteria is killing plankton and drifters There is less Fish For the humans to eat.
13. *There is an important relationship among organisms in the ocean. This relationship exists among the smallest sea life (bacteria and phytoplankton) to the largest sea life (whales). What is this relationship called and what happens to make it so important to ocean life?*
    It is so important because of bacteria died Phytoplankton would die then the animals that eat phytoplankton will die and so the Food web is The most important to the ocean life. becouse They all feed on other sea life.

FIGURE 6
Student Responses on Problem-Solving Tasks Throughout the Year

14. *Many different types of ocean research occur today. Equipment has been developed to study many aspects of the ocean from the surface to the ocean floor. If money for research became very limited, which of the following would you think is the most important to continue studying (choose one): ___minerals ___plants and animals ___ ocean floor*
I picked ocean Floor characteristic and depth because at The ocean floor you can find Tin maby even oil. People can explore and serch new plases people have never sean The guyot=Eroded seamount. They mite find other animals For sientests to study.

February Samples (Plate Tectonics)

15. *Tell how life might be different today if the continents were in different locations due to plate movement.*
If nower continent was up north it would be cold. If it went down and it was siting on The equator it would be not and they would be wering shorts and ... unlike being in The north or south. If we got colder we would eat different stuff because of how cold it is some foods mite not grow in plases unlike being at the equator. A nother way it could Be diffrent is the way peaple Travile like by water and by sledds. Different shap like if mexico broke off. The oceans would be bigger or smaller and it would change shapes.
16. *Why have volcanoes formed at spreading boundaries?*
Magma comes up and colls and makes new crust.
17. *Explain what occurs at a colliding boundary.*
The lover one go's down under The liter one.
18. *How do convection currents cause plates to move?*
Heat makes them spread.
19. *Why is plate tectonics theory a useful tool to predict future changes in the earth's crust?*
Plate tectonics can tell you what the earth looked like 2,000,000 years ago.
20. *Most earthquakes occur at or near plate boundaries. Tell why this is so.*
There is no Plate boundarie in the middle of the Plate because earthquakes occur at a Eaultor a Plate boundaries.
21. *On the planet of Kronoid an explosive device was ignited. No S waves were recorded on the far side of Kronoid. Predict what geologists would say about the interior of Kronoid. Explain why.*
The middle of the earth is not a solid it is a liquid because s waves do's not go through liquid. The senter of the planet can be magma water.

March Samples (Earthquakes)

22. *An area in the middle of a continent shows evidence of ancient earthquake activity. This place is not currently near a plate boundary. Predict what a geologist might say was the cause of that earthquake activity.*
maybe There is a under ground Falcano. That arupled a long Time ago.
23. *Which do you think might cause more destruction on earth: ___a volcano with silica magma ___a volcano with basaltic magma? Make a choice and tell why you think so.*
I think silica will do The most damage because It will make a big explosion and spred rock and ash. It is bad for people because They could get hit withe The rocks or they could Rreat in The ash.
24. *All three types of seismic waves can cause damage. Why would L waves most likely cause the most damage in a major metropolitan area?*
The L waves go on the selfes of the earth and They go slow so they cause the most damage.

May Samples (Astronomy)

25. *You are an astronomer and want to know more about the stars in a particular galaxy. What types of observations would you make to gain information about the stars?*
The first thinkg I would want to know is The size of the star or where They are and the shape. I would want to know This stuff because I am curious if I would want to get closer or to get more info. on it or not. The way I would get info. on it is with a prob. and I would look at it Through a Telescope. I want To know about it because it mite have some kind of waves or rays on it that could help us on earth.
26. *How might scientists use information obtained from studying the radiant energy that comes from objects in space?*
cars That ran By different kind of power. You could have solar heat and light.
- map locatoin of new objects in space.
- new uses for radiant energy (to help people on Earth)
- To discover effects on people or environment.
- learn how stars/planets are formed.

27. *Optical, refracting, and reflecting telescopes have been designed to observe distant objects in space. However, scientists still find it difficult to make all the observations they need. What do you think makes these observations difficult?*
mayBe because if you make a lens bigg enuf to see out Far enufe it could Bend and it would not thet The sun light go in right.

FIGURE 6
(continued)

TABLE 1
Key for Aligning Successive Tasks in Example by Intellectual Operations

Essay No.	Unit Topic	Intellectual Operation
1	Atmosphere	Explanation
2		Evaluation
3	Water Cycle	Explanation
4		Explanation
5		Prediction
6		Explanation
7	Weather	Prediction
8		Explanation
9		Prediction
10		Explanation
11		Prediction
12	Ocean	Prediction
13		Explanation
14		Evaluation
15	Plate Tectonics	Prediction
16		Explanation
17		Explanation
18		Explanation
19		Explanation
20	Earthquakes	Explanation
21		Explanation
22		Explanation
23		Evaluation
24		Explanation
25	Astronomy	Explanation
26		Prediction
27		Explanation

reliability estimates. For eight student samples an initial disagreement beyond a 1-point difference was discussed. For four units (astronomy, earthquakes, plate tectonics, and water cycle), reliability ranged from .75 to .79. The remaining three units had reliability estimates above .84.

From a criterion-referenced perspective, we can see which specific concepts and principles are being learned in any given unit. Clearly, this student is learning to solve problems in earth science. By reading individual responses, we literally can see misconceptions and appreciate his sense of understanding of how things work (e.g., a physical conception of changing states of matter). More importantly from our vantage is to take an individual-referenced perspective and ascertain whether instruction is getting him to use more concepts, use them more densely, and manipulate information with greater quality and richness.

As can be seen in Figure 7, the student produces slightly more words in his responses at the end of the year than in the beginning. He also shows increasing word production within each of the monthly units, which are reduced when a new unit is introduced. The general picture is a series of increases and decreases, with, at best, an overall slight gain. When performance is analyzed by counting the number of thought units (see Figure 8), we see no such pattern. Little change appears within units, and the overall level at the end of the year may be little different than at the beginning. At best, we see an occasional spike, which seems unrelated to time or experience within a unit.

Finally, we look at the overall quality of the response, using the flowcharts (Figure 9). In general, a small improvement is apparent throughout the year, reflecting much the same outcome as the word production measure. Within many

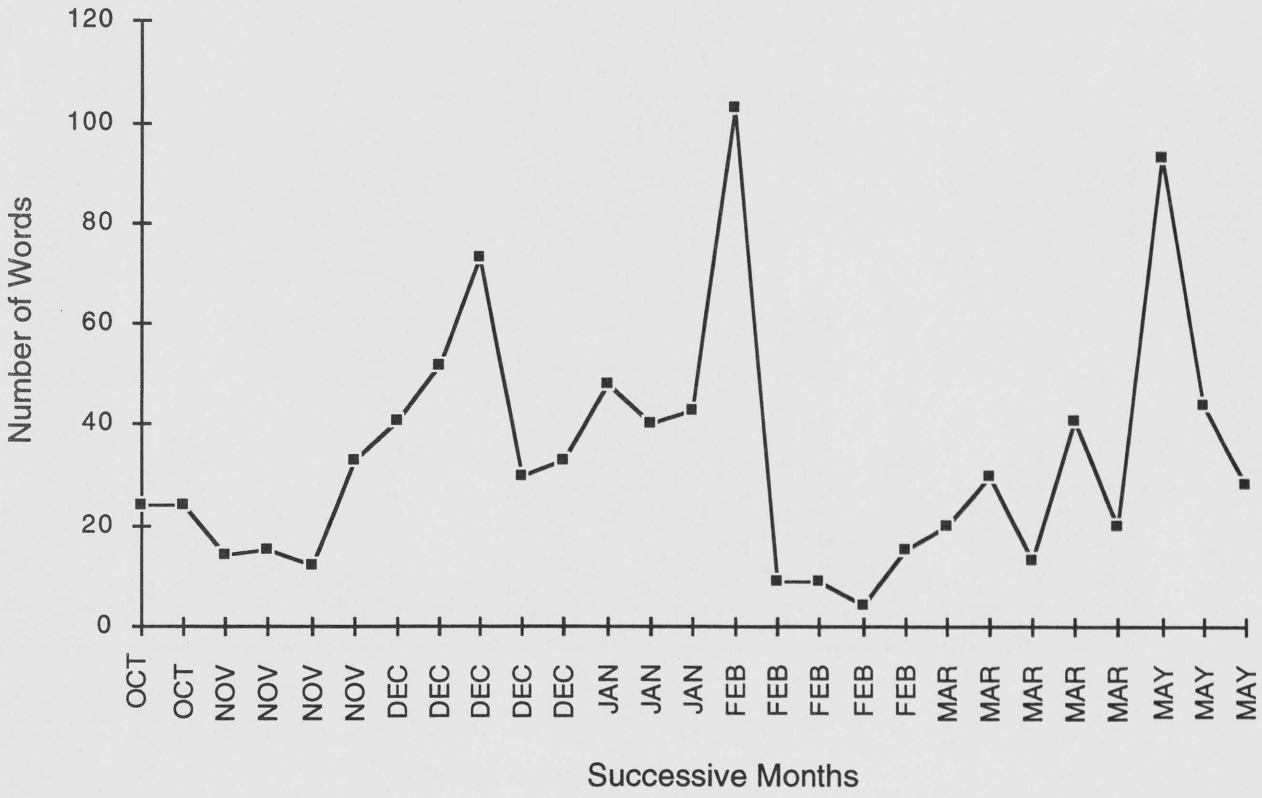


FIGURE 7
Number of Words Written in Successive Essays Throughout the Year

units, growth seems apparent, and, overall, the end-of-year performance is slightly above that attained in the beginning. Five of the last eight data points are above the first five data points. Other than the unit on plate tectonics (which has a string of low scores), most of the qualitative judgments on many of the units (weather, ocean, earthquakes, and astronomy) are above the highest value attained initially(a score of 3). Of course, they also are certainly higher than the four initial values of 2. At the same time, a split middle calculation of slope reveals an overall lack of positive trend. In conclusion, any positive outcomes must be qualified.

SUMMARY

In this article, we have focused on applying CBM to secondary content in middle and high schools. The most important dimension has been to conceptualize assessment from a repeated measurement perspective, avoiding a criterion-referenced view in which learning is considered mastery of knowledge forms. To accomplish this view, we provided a comprehensive analysis of critical thinking. Many of the features described in this article are part of the assessment tasks and are needed to provide a broad view of students manipulating information that is not bound to the content-concepts in particular but, rather, are "general-case" oriented. This view looks at whether instruction is working well or should be changed. In the end, we want to focus teaching on learning, not on curriculum or methods.

REFERENCES

Alexander, P. A., & Hare, V. C. (1989). Cognitive training: Implications for reading instruction. In J. N. Hughes & R. J. Hall (Eds.), *Handbook of cognitive behavioral approaches in educational settings* (pp. 220–246). New York: Guilford.

Alexander, P. A., & Judy, J. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research*, *58*(4), 375–404.

Alexander, P. A., Schallert, D. L., & Hare, V. C. (1991). Coming to terms: How researchers in learning and literacy talk about knowledge. *Review of Educational Research*, *61*(3), 315–343.

Anderson, L. W., & Liu, J. M. (1980, April). *The applicability of three approaches to item writing to the assessment of different types of instructional objectives.* Presented at meeting of American Educational Research Association, Boston.

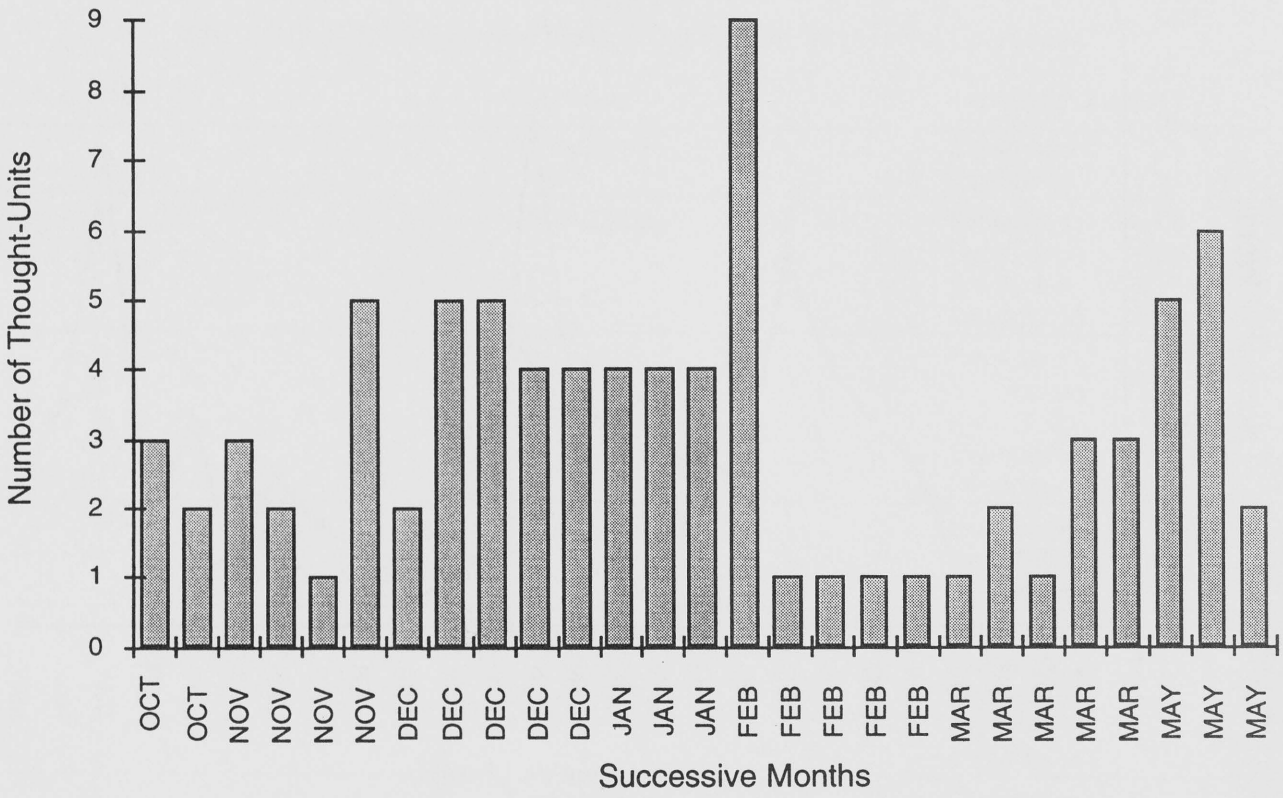


FIGURE 8
Number of Thought Units in Successive Essays Throughout the Year

Armbruster, B. B., & Anderson, T. H. (1984). Structures of explanation in history textbooks or so what if Governor Stanford missed the spike and hit the rail? *Journal of Curriculum Studies*, *16*(2), 181–194.

Armbruster, B. B., Anderson, T. H., & Ostertag, J. (1989). Teaching text structure to improve reading and writing. *Reading Teacher*, *42*, 130–137.

Baker, E. L., Ashbacher, P., Niemi, D., Chang, S., Weinstock, M., & Herl, H. (1991, April). *Validating measures of deep understanding of history*. Paper presented at the meeting of the American Educational Research Association, Chicago.

Barr, B. B., & Leyden, M. B. (1989). *Life science*. Menlo Park, CA: Addison-Wesley.

Bloom, B. S., Engelhard, D. E., Furst, J., Hill, W. H, & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: cognitive domain*. New York: Longman.

Cheney, L. V. (1987). *The American memory: A report on the humanities in the nation's public schools*. Washington, DC: National Endowment for the Humanities.

Deno, S. L. (1990). Individual differences and individual difference: The essential difference of special education. *Journal of Special Education*, *24*(2), 150–159.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, *49*(1), 36–45.

Deno, S. L., Mirkin, P. K., & Marston, D. (1980). *Relationships among simple measures of written expression and performance on standardized achievement tests* (Research Report No. 22). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Deno, S. L., Mirkin, P. K., Lowry, L., & Kuehnle, K. (1980). *Relationships among simple measures of spelling and performance on standardized achievement tests* (Research Report No. 21). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Espin, C. A., & Deno, S. L. (1993). Content-specific and general reading disabilities of secondary-level students: Identification and educational relevance. *Journal of Special Education*, *27*(3), 321–337.

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

Fuchs, L. S., & Deno, S. L. (1994). Must instructionally useful performance assessment be based in the curriculum? *Exceptional Children*, *61*(1), 15–24.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues & Practices*, *13*(3), 5–16.

Martorella, P. (1972). *Concept learning: Designs for instruction*. Scranton, PA: Intext Educational Publishers.

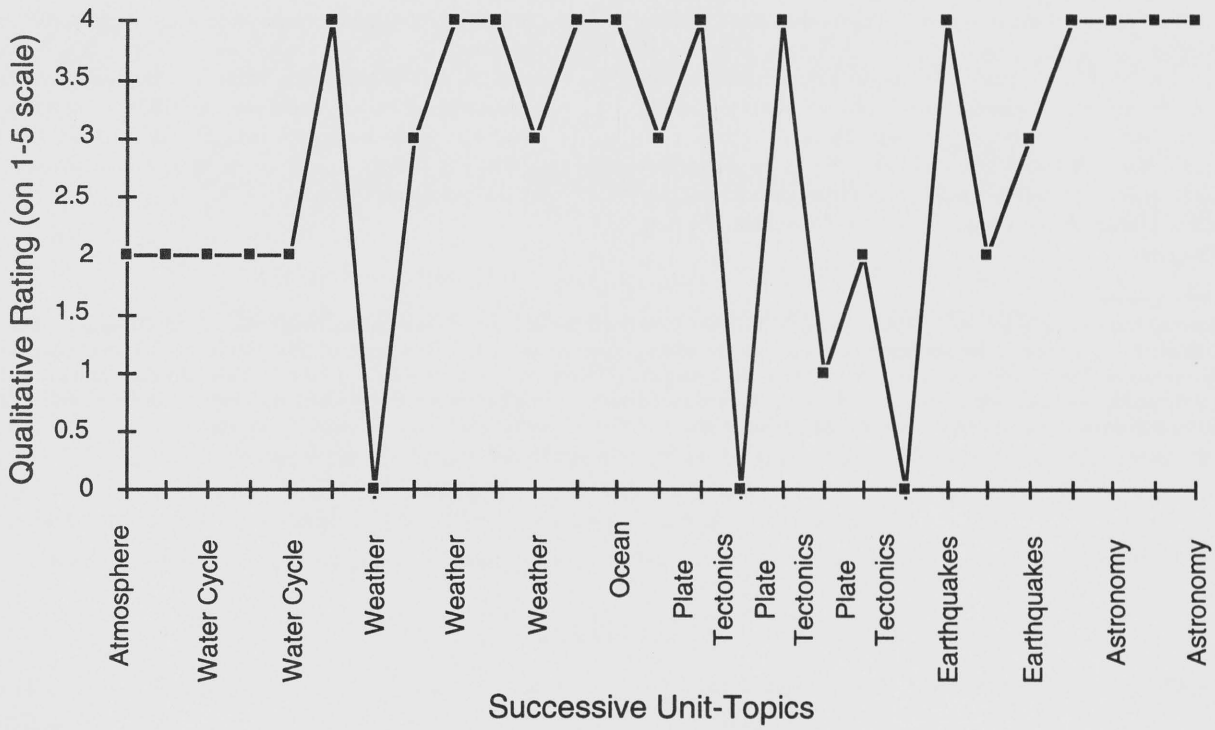


FIGURE 9
Student's Responses on Problem-Solving Tasks Throughout the Year

Mayer, R. E. (1989). Models for understanding. *Review of Educational Research*, *59*(1), 43–64.

McKenna, M. C., & Robinson, R. D. (1980). *An introduction to the Cloze procedure: An annotated bibliography*. Newark, DE: International Reading Association.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement—3rd ed.* (pp. 13–104). New York: Macmillan.

Meyer, B. J. F., Brandt, D. M., & Bluth, G. J. (1980). Use of top level structure in text: Key for reading comprehension of ninth-grade students. *Reading Research Quarterly*, *16*, 72–103.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*(3), 229–258.

Nickerson, R. S. (1985). Understanding understanding. *American Journal of Education*, *93*(2), 201–239.

Nolet, V., & Tindal, G. (1994). Curriculum-based collaboration. *Focus on Exceptional Children*, *27*(3).

Nolet, V., & Tindal, G. (1993). Special education in content area classes: Development of a model and practical procedures. *Remedial & Special Education*, *14*(1), 36–48.

Oregon Department of Education (1989–1992). *Technical report: Oregon Statewide Assessment*. Salem: Author.

Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic reader. *Contemporary Educational Psychology*, *8*, 293–316.

Prater, M. A. (1993). Teaching concepts: Procedures for the design and delivery of instruction. *Remedial & Special Education*, *14*(5), 51–62.

Quellmalz, E. S., (1985). Developing reasoning skills. In J. R. Baron & R. L. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 86–105). New York: Freeman.

Roid, G. H., & Haladyna, T. M. (1982). *A technology of test item writing*. New York: Academic Press.

Ryle, G. (1949). *The concept of the mind*. London: Hutchinson.

Seddon, G. M. (1978). Properties of Bloom's taxonomy of educational objectives for the cognitive domain. *Review of Educational Research, 48*, 303–323.

Skemp, R. R. (1978). Relational understanding and instrumental understanding. *Arithmetic Teacher*, *26*, 9–15.

Stiggins, R. J., Griswold, M. M., & Wikelund, K. R. (1989). Measuring thinking skills through classroom assessment, *Journal of Educational Measurement*, *26*(3), 233–246.

Stoltman, J. P. (1989). *Latin America and Canada*. Glenview, IL: Scott, Foresman, and Company.

Tindal, G., & Germann, G. (1991). Mainstream consultation agreements in secondary school. In G. Stoner, M. Shinn, and H. Walker (Eds.), *Interventions for achievement and behavior problems* (pp. 495–518). Washington, DC: National Association for School Psychologists.

Tindal, G., & Marston, D. (1990). *Classroom-based assessment: Evaluating instructional outcomes*. Columbus, OH: Charles Merrill.

Tindal, G., & Nolet, V. (in press). Serving students in middle school content classes: A heuristic study of critical variables linking instruction and assessment. *Journal of Special Education*.

Tindal, G., Nolet, V., & Blake, G. (1993). *Focus on assessment and learning in content classes* (Training Module No. 4). Eugene: University of Oregon Resource Consultant Training Program.

Valencia, S., Pearson, P. D., & Chapman, C. (undated). *New strategies for reading comprehension assessment: Illinois initiatives*.

Williams, R. G., & Haladyna, T. M. (1982). Logical operations for generating intended questions (LOGIQ): A typology for higher level test items (pp. 161–186). In G. H. Roid & T. M. Haladyna (Eds.), *A technology of test item writing*. New York: Academic Press.

This project was funded by a special projects grant (H029K10130) from the Office of Special Education Programs, United States Department of Education. However, the opinions expressed herein do not necessarily reflect the position of the U.S. Department of Education or the College of Education at the University of Oregon, and no official endorsement by the Department, College, or University should be inferred. Two individuals provided critical help in preparing this manuscript. Sandra McCollum was responsible for working with the student and the content teachers over the year, and Dewayne Joehnk was instrumental in helping develop the scoring flowcharts.

DOWN SYNDROME: Birth to Adulthood

Giving Families an EDGE

John E. Rynders
University of Minnesota

J. Margaret Horrobin
Minneapolis Pediatrician

Filled with photographs of family activities, *Down Syndrome: Birth to Adulthood* is an invaluable guide for parents of children with Down syndrome, as well as the professionals who help advance whole-family development. The authors, well-known and respected leaders in the field, use case studies of families from the EDGE Project at the University of Minnesota to examine all the issues—both large and small—in raising a child with Down syndrome from infancy to adulthood. Research findings from several sources are woven into the pages in a clear, readable fashion, followed with numerous practical suggestions.

All royalties from this book's sale are donated to nonprofit organizations serving people with Down syndrome and their families.

Contents

9502*/paperback/ISBN 0-89108-236-0/***$29.95**

Love Publishing Company
1777 S Bellaire Street
Denver, CO 80222
303-757-2579 • 303-782-5683 (FAX)

professional update

April 5–9, 1995

CEC Annual Conference
Indiana Convention Center, Indianapolis

Contact: Council for Exceptional Children
1920 Association Drive
Reston, VA 22091

April 10–13, 1995

Fourth International Special Education Congress
Birmingham, England

Contact: John Visser
School of Education
University of Birmingham
Birmingham B15 2TT
England
FAX 021-414-4865

April 18–22, 1995

American Educational Research Association
San Francisco Hilton Hotel

Contact: American Educational Research Association
1230 17th Street, NW
Washington, DC 20036