GOALS 2000 Implementation:

Assessment Definitions

Analytical trait scoring: a performance is judged several times along different important dimensions or traits. (We're *analyzing* a performance along various *traits*.) For example, we could judge student performance on an open-ended math problem for the traits of: understanding the problem, correct use of procedures and strategies, and the ability to communicate clearly what was done. Use of a scoring rubric and anchor papers for each trait is common, but, you could also have a checklist. (Analytical trait scoring can be generalized or task specific.)

Benchmark Performance: examples of papers or performances that serve as a standard against which other papers or performances are compared; often used as examples of performances at different points on a scoring rubric. The best the performance or product can be done or has been done which anchors the top point on a scoring scale.

Benchmark is also used to describe

a. expected or anticipated skill or understanding at various developmental levels
b. critical learning points

Benchmarks and estimation: numbers that are easy to work with and picture as well as non-standard units such as familiar objects or events that a student can use as referents for estimating.

Constructed response (task): generating a response to a question rather than choosing an answer from a list; for example, short answer, extended responses, and portfolios.

Extraneous interference (also called "mismeasurement, bias, and distortion")**:** features of an assessment task or context that might mask student achievement rather than reveal it - for example, excessive reading on a mathematics problem solving test, or role-playing on a science assessment. What is actually assessed might be reading or role-playing rather than problem solving or science skills.

Generalized scoring: a rubric that can be used to score performance on a large number of related tasks - for example, a writing rubric that can be used for most writing assignments or a math rubric that can be used for most open-ended math problems. The approach can be holistic or analytical trait. The format can be a checklist or a rubric.

Growth: measured from where one starts.

Holistic scoring: a single, overall score is assigned to a performance. The format can be a checklist or a rubric. The approach can be generalized or task-specific.

[Source: Improving Classroom Assessment: A Toolkit for Professional Developers, Regional Educational Laboratories, February 1998]

Indicator: a specific description of an outcome in terms of an observable and assessable behavior. An indicator specifies what a person who possesses the qualities articulated in an outcome understands or can do. For example, indicators of good critical thinking might be: considers multiple perspectives; bases decisions on reputable sources of information; and looks for extraordinary proofs for extraordinary claims. Indicators are often used in rubrics to cue raters to the right score.

Ill-structured (tasks or questions). *A question is "ill-structured" if there is no recipe to answer it, if it is non-routine.* Such questions thus demand more than knowledge: they demand good judgment and imagination. All good essay questions or design problems are thus ill-structured: even when you know what is expected you have to scratch your head and methodically craft and refine -- often revising as you go -- your solution or product. Often with such problems, the student fully understands the intended result, but may not know how to solve the problem or get productively under way, even with a great deal of prior training.

Most "authentic" problems are ill-structured; most test "items" are not. Test questions are typically "well-structured" in the sense that there is a single, unambiguous right answer -- or an algorithm for solving it; it is neat and clean. Such "items" *are fine for validly assessing elements of knowledge but not appropriate for judging the student's ability to use knowledge wisely -- namely, how to judge which knowledge and skill to use when*. (Think, for example, of the differences between the "test" of each drill in basketball, and the "test" of playing the game well in performance.)

Note, however, that ill-structured problems are not the same as *open-ended* problems (see below). *All open-ended questions are ill-structured; the reverse is not necessarily the case.* Ill-structured problems can have clear, unambiguous and correct answers (as in the task of designing the cheapest trip to Europe), but there still is no mechanical procedure for solving the problem correctly.

On-demand assessment: assessment that takes place at a predetermined time and place. State tests, SATs, and most final exams are examples of on-demand assessments.

Open-ended tasks: tasks that require students to generate a solution to a problem, but there is no single correct answer. It requires non-routine thinking-strategy and good judgment. Rather, there is a range of appropriate responses that can be obtained by a variety of solution paths. An example is: "Below you see a bar graph without any labels. What might this be a graph of?"

Open-response tasks: the kind of performance required of students when they are required to generate an answer, rather than select it from among several possible answers, but there is still a single, correct response. An example is: "There are four pieces of wood, each measuring seven feet. If you used them as a fence around your square yard, how large an area would you create?"

Performance assessment: direct, systematic observation of actual student performances and judging the quality of the performances according to pre-established performance criteria.

[Source: Improving Classroom Assessment: A Toolkit for Professional Developers, Regional Educational Laboratories, February 1998]

Performance criteria: a description of the characteristics on which the response to a task will be judged. For example an oral presentation might be judged on content, organization, delivery and use of language; and here's what we mean by delivery,…Performance criteria may be holistic, analytical trait, general or task-specific. Performance criteria are expressed as a rubric or scoring guide. Anchor papers or benchmark performances may be used to identify each level of competency in the rubric or scoring guide.

Primary trait scoring: a scoring procedure by which products or performances are evaluated by limiting attention to a single criterion or a few selected criteria. These criteria are typically based upon the trait or traits that are most essential to a good performance. For example, if a student is asked to write to the Department of Energy urging the opening or closing of a nuclear power plant, the primary traits might be the ability to communicate persuasively and the correct application of scientific knowledge to back up one's position. Scorers would attend only to these two traits, and not, for example to spelling.

Rubric: an established and written-down set of criteria for scoring or rating students' performance on essays, portfolios, or other performance tasks.

Selected-response assessments: students select the correct response from among a set of responses offered by the developer of the assessment. Multiple-choice and matching tests are examples of selected-response assessments.

Standardized assessments: assessments that are administered and scored in exactly the same way for all students. Traditional standardized tests are typically multiple-choice, norm-referenced, machine-scored, and designed to measure skills and knowledge that are thought to be taught to all students in a fairly standardized way. But, performance assessments and other criterion-referenced tests, can also be standardized if they are administered and scored in the same way for all students. Standardization is an important consideration if comparisons are to be made between scores of different individuals or groups.

Task (as in a "performance task"): an assessment exercise - for example, a particular math problem to solve, a lab to do, or a paper to write. Performance tasks don't have to be stand-alone activities that occur at the end of instruction. For example, students working in groups on a project provide a naturally-occurring task on which to observe group-process skills; or students reading aloud provide a naturally occurring way to assess reading fluency.

Task-specific scoring: a scoring guide or rubric that can only be used with a single exercise or performance task. For example, "to get a '4' students must show a picture of a whole divided into four equal parts." A new rubric is developed for each task. The format can be a checklist or rubric. The approach can be holistic or analytical trait.

[Source: Improving Classroom Assessment: A Toolkit for Professional Developers, Regional Educational Laboratories, February 1998]