9 Curriculum Evaluation

LEARNING OUTCOMES

After reading this chapter, you should be able to

1. Discuss the nature and purpose of evaluation
2. Articulate the assumptions behind the scientific, modernist approach and the humanistic, postmodernist approach to evaluation
3. Explain the objectives of the scientific, modernist evaluation model and the humanist, postmodern evaluation model
4. Distinguish between high-stakes, norm-referenced, and criterion-referenced tests and explain the rationale for employing each type
5. Describe various alternative assessment strategies
6. List the various issues regarding fairness in curriculum evaluation

People agree that curriculum evaluation is essential to curriculum development, implementation, and maintenance. However, they disagree regarding evaluation's meaning and purposes, how to approach evaluation, and how to employ its results. Ideally, evaluation determines the value of some action or program, the degree to which it helps students meet standards, and its importance. Implicitly and explicitly, evaluation reflects value judgments about previous curricula and instructional designs. Evaluation critiques previous documents, plans, and actions.

We define *evaluation* as synonymous with *assessment.* We believe that assessment (evaluation) involves value judgments as to merit and worth. These judgments affect which data we gather and how we view those data. Evaluation requires of educators actions to judge the appropriateness of both their and students' actions. In evaluating students' learning, educators often give tests that assess what educators consider important. Teachers critique the quality of their teaching often by viewing videotapes of an instructional session.

In curriculum evaluation, attention focuses on both teachers' and students' actions that result in students' learning specific contents and skills. Today, curriculum evaluation is more challenging than in the past. Currently, education in general, and schools in particular, exist in a dynamic complex in which social, economic, political, and technological changes generate diverse views as to the school's purposes and the intellectual competencies and skills that will serve students well. As Peter M. Taubman asserts, we are living in a transformative time under the "twin banners

of 'standards' and 'accountability.'"[1] Standards and accountability are battle cries often uttered by noneducators, particularly politicians and business leaders, with no idea as to the nature of curriculum and instruction. Most members of the public lack backgrounds in psychometrics and are especially unaware that holding educators accountable for attaining standards and also addressing diversity of students and the need for creativity in schools are often at cross-purposes.[2]

Certainly, educators should have standards and be accountable. But what does that mean? Taubman articulates that we are being consumed by an "audit culture" in which educational programs, practices, and discourses are being encapsulated, standardized, and reduced to sterile quantifications. We seem to be functioning under a cloud of doom. To avert this doom, many are urging an enactment of a one-size-fits-all program and performance. All students must learn particular subject matter and must demonstrate identical proficiency.

Much of this cloud of doom is enhanced by the myth that when compared to other students in developed countries, our students do not measure up as having comparable competencies. We are not number one! But, few challenge the notion that we must be number one, first, best in the world community. And can one determine competence, understanding, creativity, and general knowledge in cognitive, affective, and psychomotor realms by taking a standardized test? What does a score really indicate? We suggest the only precise statement that one can deduce from a test score is that someone or a group of someones attained a score—a number—higher or lower than we did. Can we state that a high score means that students actually know more than one with a lesser score? And how can we define what "knowing more" actually means?

Further, can we really utilize a test score or cluster of test scores to indicate that our schools are failing, that they need to improve? As David Berliner and Richard Glass denote, judging and critiquing school systems in our industrial world is no easy task. School systems exist within national contexts. In our heterogeneous country, it is misleading and perhaps dangerous to make general assertions about the effectiveness of our curricula and instructional strategies on students' learnings, skills, competencies, and dispositions.[3] And, as Berliner and Glass posit, how do we define better? They point out that even if we can define better, we need to query, "better for whom? better on what criteria?"[4] We also raise queries as to what "better" means when considering students' creativity, tolerance, empathy, risk taking, social skills, poise, and adventurousness.

When engaged in education and assessing the benefits of that education, we must denounce the notion that students are widgets that are to be standardized and measured accordingly. This is an underlying assumption that has been part of American culture since the 1800s. It assumes that schools are factories and students are products that populate assembly lines. In 1916, Stanford University professor Ellwood Cubberley declared the following:

> Our schools are, in a sense, factories in which the raw products (children) are to be shaped and fashioned into products to meet the various demands of life. The specifications for manufacturing come from the demands of twentieth-century civilization, and it is the business of school to build its pupils according to specifications laid down.[5]

We vehemently refute the notion that schools are or should be factories. We also challenge the belief that schools and their curricula and instructional strategies are to be determined by industry. Schools are designed to produce educated citizens. Schools are not places for training people for specific jobs in industry. Yet, in the 21st century, we have a wave of individuals from industry, government, society, and even some educators urging STEM for all: science, technology, engineering, and mathematics. The December 2014–January 2015 edition of *Educational Leadership*, a publication of ASCD, was so titled.[6] Why? We need more engineers; we need more technology experts; we need more scientists; we need more mathematicians. Why such a need? We are behind China in the numbers of engineers being graduated. We must be first. No one seems distraught that we are not producing greater numbers of literary figures, artists, political scientists, or historians. Of course, some individuals are urging that we add to STEM, such as the arts and the humanities. But, arts and humanities do not build computers or airplanes, or

furnish new chemicals and 21st century technologies. Those in industry are still recommending that schools continue as factories that produce widgets designed to industry specifications.

Curriculum evaluation should not exist to give us bragging rights in the world community. Certainly, evaluation is a necessary activity to access how our curriculum and instruction are addressing and challenging the educational development, writ large, of our students. But, as Berliner and Glass point out, educational assessment cannot be equated with the Olympics. International tests are not an educational Olympic event. These authors point out that "equating national rankings on international standardized tests with rankings in athletic events is simply deeply flawed logic."[7] We assert that employing such logic does not give an accurate understanding of the quality of American education. Nor does it present data that provide insight as to the nation's economic and educational health.

So, why do we as a society have this national perception that we are falling behind, that our schools are failing us? Partly what is driving this chaos regarding curriculum evaluation is the No Child Left Behind Act, signed into law in January 2002 by President George W. Bush. This law is a directive that all educational agencies, schools included, at national, state, and local levels will work to create and then evaluate educational programs. It articulates that states will determine academic standards at three levels of achievement: basic, proficient, and advanced. However, it notes that 100 percent of students must be proficient on state and reading standards as determined by state-created examinations in reading and mathematics. Proficiency must be attained in science. Such must be accomplished if the United States is to be competitive in the world.[8]

Proficiency suggests high standards, but we must ask, "How high and for whom?"[9] What about just achieving the standard at the basic level? How do we report standards attained at the advanced levels? Must everyone engage in the same behavior? And must this all be evaluated by high-stakes exams administered at specific times? Also, can we say with certainty that a high score on a mathematics exam translates into student success 10 years later? Will achieving a high standard of history knowledge mean that a student will be an effectively contributing citizen 15 years hence? What about the uniqueness of individual students? And can everything that we wish to accomplish be measured by an exam? How does one measure empathy and tolerance? What is basic empathy? How does one measure proficiency in empathy?

Certainly, there are ways to measure students' attainments in knowledge and action other than employing high-stakes texts. However, No Child Left Behind (NCLB) seems to celebrate standardized tests as the primary means of gathering data to determine schools' accountability. Every state must build an accountability system that utilizes tests that validly measure student learnings, levels of achievement, and teacher effectiveness. Results from these tests must be disaggregated to take into account "socio-economic status, gender, race, ethnicity, disabilities, and levels of English language proficiency."[10] This directive seems to contradict that 100 percent of students will be proficient. And, if we take into consideration students with limited English-speaking skills or students with learning disabilities, then we cannot simultaneously have 100 percent student proficiency.

Although schools have been ordered to develop curricula and evaluation means to document that no child is being left behind, the order does not indicate how the states are to develop such tests. There is no nationwide testing policy. Also, there does not seem to be much guidance from the federal Department of Education as to how to address the unique cultures and subcultures within their states. New York State, which has had the New York State Regents exam in place long before NCLB, is certainly culturally different from New Hampshire or New Mexico.

Further complicating curriculum evaluation are the explosions of knowledge regarding how the brain functions, how people learn, how the political realm affects schooling, how new pedagogies can address the needs of diverse student populations, how curricula can be created using various modern and postmodern approaches, and how assessment devices can be created and modified to get at the essences of learning. Educators should use evaluation methods and approaches that draw on the latest thinking. Yet, in some ways, the tests we currently use are based on 19th century psychology.

James Pellegrino, Naomi Chudowsky, and Robert Glaser note that our current approaches to evaluation do not adequately take into account the increase in knowledge about how the brain functions. We already measure students' learning processes and knowledge of basic facts, and we derive estimates of students' command of particular curriculum areas, but we fail to get an accurate picture of the depth and breadth of students' knowledge and cognition. Current evaluation approaches do not provide views of the complex knowledge and skills required for learning.[11] They do not adequately address student creativity, compassion, commitment to action, or enthusiasm.[12]

Current evaluation takes snapshots of student achievement with regard to knowledge and process at particular points in time. Washington State obtains data on students' achievement at grades 4, 7, and 10, but not a view of how, for example, students' understandings and skills evolve. Testing students three times a year shows that scores are going up, going down, or remaining level, but it does not necessarily indicate the amount of learning.[13]

Adding to the difficulty of evaluating the curriculum is the increasingly voiced demand that assessment be fair and appropriate for diverse students. Certain segments of society express concerns that tests, especially standardized ones, favor certain student populations. Others argue that standardized tests are not fair because they focus on subjects, topics, and processes that have not been taught in their schools. Also, some claim that the standards set for passing these tests harm less-advantaged students.

A real challenge in employing standardized tests to measure the quality of curriculum and the effectiveness of instructional strategies lies within the nature of test design itself. Wayne Au illuminates a problem with test design using the Scholastic Assessment Test (SAT) as an example. The Educational Testing Service (ETS), which primarily oversees the management of students taking this test, employs one of the six major divisions of the test as an experimental section to ascertain what questions might be potentially valid to include in future SATs. Data from this "trial" section of the test furnish psychometricians with information that allows them to either keep a question or discard a question.

The psychometricians determine the difficulty levels of these questions and how various groups of students have done in the past, specifically White students, students of color, male students, and female students. The test designers have a database that historically indicates that Whites outperform students of color and other minorities. This favoring of Whites indicates that the questions are valid when the experimental questions in the SAT reflect those results. However, if students of color outperform Whites on particular questions by a significant margin, then the test makers usually classify the questions as psychometrically invalid. They are then excluded from future SATs.[14] How to deal with this validity factor is a real challenge for those using standardized tests in this century with an increasingly diverse student population.

Today, with regard to curriculum evaluation, we are not only judging whether students are learning the curriculum effectively, but we are also charting teachers' instructional competence. We assume that the curriculum developed and implemented is of value and is worth knowing. Doubting this, we would not teach it. The value of the curriculum developed and presented is a given. However, in the current evaluation climate, some are suggesting that teachers' pay be connected to how well they teach the curriculum. Effective teaching translates into high test scores. Effective teachers will receive higher salaries. Some even have suggested that competent teachers of high-status subjects such as mathematics and science receive larger salaries because their subjects are more crucial to the nation's welfare. Such a notion would violate a rule of merit pay: create a program that encourages collaboration.[15] It likely will foster a deleterious competition among school faculty.

There is much dialogue centered on evaluating and rewarding teacher performance with merit pay. There are several myths regarding merit pay, as articulated by Chris Hulleman and Kenneth Barron.[16] The first myth is that performance pay systems improve performance. Performance pay enhancement may increase performance quantity, but that does not necessarily equate with quality. Students might learn more, but their understanding may not be increased essentially.

A second myth advanced by Hulleman and Barron is that performance pay systems will heighten teacher motivation. They cite research that indicates the opposite; expected rewards based on performing a task at a specific level actually undermine intrinsic motivation for performing the task. Most teachers do not engage a particular instructional strategy to gain a pay raise. The authors also point out the danger of applying motivational business strategies to the educational arena. Additionally, although a business person can document an increase in sales to request a bonus, it is far more difficult to quantify quality learning by students.[17] Results of quality teaching may not appear for decades. Should we give delayed bonuses if students in later years create a new business or win a Nobel Prize? How can we make an evaluation of a teacher's impact on a student? The challenge is to make a direct causal connection from a teacher's action and a student's future accomplishment. It cannot be done.

Despite our protest, the idea of teacher pay tied to performance is not going to disappear. Quite likely, it will increase as a clarion call for improving schools. We are not going to separate curriculum evaluation from teacher effectiveness in "teaching" the curriculum that is delivered. As Matthew Springer and Catherine Gardner note, we are entering a perfect storm: Teacher compensation is being battered by performance and market-oriented pay policies.[18]

Although we assert that we cannot make a direct causal connection between a teacher's action and student's knowledge or skills, this argument, according to Springer and Gardner, may be losing validity. They note that many states and school districts have created sophisticated longitudinal data systems that enable determining links between individual student performances and teachers' instructional strategies. They note that with such data systems, one can more precisely estimate teachers' contributions to students' learnings. Additionally, they note that there is increasing research that aims to develop and validate sophisticated measures of effective teaching.[19]

Currently most states have instituted one of these "sophisticated measures" identified as value-added measurement (VAM). The method employs processes of statistically measuring teachers' performances based on students' achievement determined by their test scores. Students' academic growth is measured each year to determine a gain in knowledge and understanding. Advocates of the VAM note they are controlling for variables outside of teacher influences: social class standing, ethnic groups, English proficiency, and learning abilities and disabilities.[20]

It does appear that the enactment of No Child Left Behind in 2002 added fuel to this move to hold schools and teachers accountable for enhanced learning and achievement. President Obama's Race to the Top fanned the flames. Even the title Race to the Top implies that schools are competitors. Berliner and Glass indicate that advocates for having schools compete and providing merit pay to teachers will stimulate increased teacher effectiveness and creativity and enhanced student learnings. They argue that more effective teachers, those whose students have higher test scores, will earn more money. This will motivate other instructors to be more effective in their pedagogies, to increase their salaries. Of course, as already indicated, it is a myth that performance pay systems will actually motivate teachers.

Education is not a sport; it is not a game with winners and losers. As indicated previously, education is not an Olympic activity. In sports, there are winners and losers. Do we want that in education? Some might say this is already the situation with American education. But, we believe strongly that making education into a competitive sport within the country and the world will further erode our educational system. And having educational competitions will not furnish us with quality education. We will know only the participants' "scores" in the educational games. We will not have any understandings of the quality of the myriad learnings and dispositions that enable individuals to excel at the game. As Berliner and Glass denote, student achievement is impacted by forces outside of the school.[21]

Despite the commentary in the previous paragraphs, we still consider evaluation essential to the continual usage of meaningful curriculum. If teachers and the community are to support the curriculum, educators must conceive and implement effective evaluation and reporting processes. They also must query their assumptions about evaluation and whether their dispositions regarding evaluation are in the modern or postmodern camp.

9.1 Value-Added Measures Explained
This video explains how Ohio plans to implement value-added measures to determine if teachers are doing a good job. What are some reasons teachers may be against value-added measures and merit pay?

https://www.youtube.com/watch?v=925RnyfzbjU

THE NATURE AND PURPOSE OF EVALUATION

Evaluators gather and interpret data to determine whether to accept, change, or eliminate aspects of the curriculum, such as particular textbooks. Curriculum evaluation is necessary not only at the end of a program or school year, but also at various points throughout the program's development and implementation.

At the beginning of curriculum development, the very concept of the program must be evaluated. Does the program have worth and merit? Throughout the process, educators must evaluate the worth and merit of the curriculum's content and experiences. Curriculum evaluation focuses on whether the curriculum is producing the desired results. For example, does it get students to perform at the level of standards indicated for student success? Evaluation identifies the curriculum's strengths and weaknesses before and after implementation. Evaluation also enables educators to compare different programs in terms of effectiveness. People want to know how their students measure up against other students at the local, state, national, and international levels.

Pellegrino, Chudowsky, and Glaser view assessment as a process of reasoning from evidence.[22] The first question in this process is, "Evidence about what?"[23] Data interpretation is possible only when we understand what we are attempting to do and know what standards we want students to meet.

The process of reasoning from evidence in curriculum evaluation can be conceptualized as an hourglass. This schema is an expansion of Pellegrino, Chudowsky, and Glaser's reasoning assessment triangle, which had the following features: cognition at the top of the triangle, and observation and interpretation at the corners of the triangle base.[24] We have added to their model, placing curriculum at one corner of the top of the hourglass with cognition at the other corner (Figure 9.1). The neck of the hourglass represents the observation stage of reasoning. The base of the hourglass represents interpretation.

The curriculum organizes subject matter in terms of scope and sequence. In curriculum development, educators must make evaluative judgments regarding the worth of the subject matter being considered and organized as well as the political and social climates within which the curriculum will exist. Educators consider this question: What evidence suggests that the curriculum contemplated, planned, and then delivered has value, meets students' and society's needs, and is consistent with curriculum theory?

Cognitive theories inform us in our data gathering. How do students acquire knowledge, construct meaning, and develop competence? Cognitive models of teaching can assist teachers in shaping their instructional approaches and evaluating students' learning.

Observation includes all the means by which data are gathered. It may involve written tests, reviews of students' work (e.g., their portfolios), and viewing students as they engage in

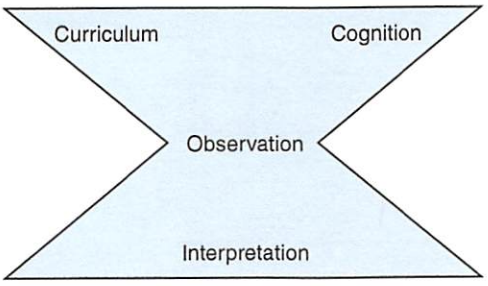

FIGURE 9.1 Process of Reasoning from Evidence in Curriculum Evaluation

Source: Adapted from James W. Pellegrino, Naomi Chudowsky, and Robert Glaser, eds., *Knowing What Students Know: The Science and Design of Educational Assessment* (Washington, DC: National Academy Press, 2001).

particular educational actions. Observation includes questionnaires, checklists, inventories, interview schedules, and video performances. It also includes data on teachers—for example, from observations of teachers, viewing of videotaped lessons, analyses of lesson plans, and interviews.

In the *interpretation* stage of curriculum evaluation, educators draw on their assumptions about curriculum and cognition. They process data into evidence regarding the curriculum's success. At the classroom level, interpretation tends to be informal and qualitative, including interpretation regarding teachers' instructional approaches. At the district level, interpretation tends to be more formal, but it still can be qualitative rather than quantitative (employing a statistical model).[25] Often, district-level interpretation is both qualitative and quantitative. Interpretation implicitly draws on theories of testing, statistical models of data analysis, and theories of decision making.

Evaluation must remain connected to the totality of curricular activities. Evaluators first must ask themselves what aspect(s) of the curriculum they wish to evaluate and what types of learning will receive focus. They then must determine which means of data gathering best suit one or more particular goals of the curriculum. Which questions will furnish the data desired for interpretation?[26]

Often, evaluators investigate the appropriateness of a particular assessment procedure or form of assessment. Frequently, evaluation centers on how to modify the staff's in-service education. Sometimes evaluation focuses on just how educators can communicate with and educate the community. Sometimes, evaluation focuses on the effectiveness of a school environment. However, most evaluation focuses on curriculum or instruction.

Evaluation that focuses on curriculum actually commences in the initial stage of curriculum conceptualization. Specifically, educators query whether the curricular content and experiences initially considered are worthy of the effort. The evaluative questions essentially reflect Spencer's question: What knowledge is of worth? And, we would add, what knowledge is of worth to the greatest number of diverse students in the 21st century? While this question might appear easily answered, to respond to it in this dynamically changing century is a Herculean task. How does one evaluate the worth of a particular content or educational experience with regard to unknown and emerging situations and demands? What content is relevant for understanding contents and situations not yet existing? For those who wish students to master contents such as the STEM subjects, what aspects of these subject contents are to be judged worthy of attention if our views of the future have not been articulated?

We assert that evaluation, or assessment, after the curriculum has been implemented focuses on two domains of activity: teacher instructional strategies and students' learning strategies. But, in the chaotic dynamics of this century, are the instructional strategies and students' learning strategies fluid enough to adjust to the dynamics of exploding knowledge realms and information technologies? What might be judged effective in 2015 may be deficient in 2020.

Catherine Taylor and Susan Nolen note that teachers first engage in assessment to gather information about students' understandings and skills.[27] Such information is gathered via various procedures so that teachers can decide what to teach and the manner of teaching and student engagement. Essentially, this is the view that assessment is a process of reasoning from evidence. They must determine the individuals' degrees of success in processing particular content and concepts. This assessment is used primarily at the commencement of a lesson or unit of study. At or toward the end of a lesson or unit, teachers map out assessment procedures to record students' mastery of some content or expertise in some skill or intellectual process. Here teachers primarily engage in observation and interpretation phases of evidence gathering. Common methods employed are tests, with teachers often assigning grades. Taylor and Nolen suggest that the final purpose of assessment is to make comparisons of their students with others; that is, to measure their students' standings compared with other students. More is said about this last purpose later in this chapter.

It does seem that evaluation, or assessment, can and does have two purposes. Lisa Carter suggests that one view is that evaluation is activated so educators can sort and select not only

curricular content and instructional strategies, but also which students experience various curricula and instructional experiences. Heavy emphasis is on employing test scores to sort and track students—that is, to place them in similar groups according to abilities, interests, and accomplishments. The second purpose of evaluation, Carter indicates, is to gather information, or evidence, in order to make educational, curricular, and instructional decisions that enhance students' learning of the curriculum being taught. Here evaluation aims to adapt the curriculum to the students rather than to mold the student to fit the curriculum.[28]

To be successful in carrying out evaluation, with emphasis on the second purpose, there are key questions to be raised. These questions, developed by Harriet Talmage in the mid-1980s, are still relevant today.

Evaluation Questions

Talmage posed five types of questions that educators can consider when evaluating curricula: questions of intrinsic, instrumental, comparative, idealization, and decision value.[29]

The question of intrinsic value addresses the curriculum's goodness and appropriateness. It deals with both the planned curriculum and the finished (delivered) curriculum. For example, a school would ask if a new language arts curriculum incorporates the best thinking to date on language arts content and that content's arrangement and presentation. Would specialists in linguistics, composition, grammar, and communication give the planned curriculum high marks? Raising such questions is not a simple matter of getting experts to analyze the curriculum document. People bring their philosophical and psychological views to the question of intrinsic value. They perceive the curriculum in light of the purpose of education that they see as paramount. (Should we stress critical thinking, citizenship, or preparation for employment?) They also see curriculum in terms of their preferred learning theory. (Behaviorists, cognitivists, and humanists have different views about content and presentation methods.)

The question of instrumental value asks, what is the curriculum good for, and who is its intended audience? Educators deal with the first part of this question by attempting to link the planned curriculum with the program's stated goals and objectives. The question of instrumental value also addresses which students accomplish what is planned in the curriculum and to what extent. The level of attainment relates to standards that reflect value preferences. Evaluation efforts should identify the types of students who are likely to benefit the most from the planned curriculum.

People faced with possible new programs often ask the question of comparative value. Is the proposed new program better than the one it is supposed to replace? Usually, new programs are created because people feel that the existing program is inadequate. When comparing programs, remember that different programs may have different goals. Is a program that stresses skills better than one that stresses contemporary world issues? Certainly, the two are different. Whether one is better than the other relates to educators' values and priorities. However, if a suggested program is of the same type as the existing program, evaluators should consider comparative value not only in terms of student achievement, but also in terms of the two programs' ease of delivery, cost, demand on resources, role in the existing school organization, and responsiveness to the community expectations.

The question of comparative value is often raised when comparing the achievement of students in various countries, if not the curricula of the countries. Currently, voices in our national educational discussion suggest that when tested, American students do not compare favorably with students in other countries. It is often noted that the standings of American students, especially in mathematics and science, do not compare well. Usually, in such comparisons we essentially are not interested in what these various students actually know. We are more attentive to just how our students compare with others. We employ such data to rank students and to determine differences between students.[30] Basing the quality of our students' achievement in mathematics or science solely on a test number provides us with scant evaluative information. It denies us evidence essential for making evaluative decisions.

The question of idealization value addresses ways to improve a curriculum. Evaluators should not be concerned only with determining whether what was planned actually happened; they should also view data in terms of ways to create and maintain the best possible programs. They consider information on how the program is working and ask themselves if there are alternative ways to make the program even better—for example, to heighten student achievement or involve students more fully in their learning. The question of idealization value should be asked throughout the delivery of the new program. Educators must continually reconsider how they might fine-tune the program's content, materials, methods, and so on, so that students will optimally benefit.

The idealization question currently is reshaped into the question of curricular and instructional improvement. This question, newly redefined, requires "finer-grained measures for detecting improvement."[31] Assessing improvements in students' performances or even changes in teacher's strategies is much like measuring the movement of a glacier. Spend the day observing a glacier, and it appears stationary. However, if you take monthly observations, you can observe whether it is advancing or retreating. Certainly, a yearly observation schedule would document movement.

In raising the idealization question, the improvement question, one should remember Fullan's comment that "changes in student performance lag behind changes in the quality of instructional practice."[32] He suggests that we need more refined assessment instruments to detect changes in students' classroom learnings and behaviors. If we neglect field studies and employ only year-end tests, we will be able to report only students' particular level of learning. We will have violated the "evidence factor" because we will be unable to articulate the little daily learnings that assisted or sabotaged student progress. The idealization question requires frequent measurement of teacher and student action, employing a variety of evaluation procedures and materials.[33]

The question of decision value deals with the vital role that the previous four questions play in the evaluation process. If those four questions have been addressed, the decisions made should be quality decisions. The evaluator and the curriculum decision-maker should now have evidence documented in such a manner that they can decide whether to retain, modify, or discard the new program. However, the question of decision value is ongoing. The value of the decisions made to date must be assessed as the curriculum is delivered in classrooms.

That the decision value question is ongoing essentially means that the previous four questions are constantly considered. Evaluation is never completed. Evaluation is challenging work. We suggest that the results we obtain and the evidence we gather are more like impressionistic paintings rather than designs generated by algorithms in a computer program. Individuals viewing an impressionistic painting draw a multitude of learnings and insights and ever-differing emotions; we must consider students more as paintings than as computer programs.

Definitions of Evaluation

Evaluation is a process whereby people gather data in order to make decisions. Apart from that generality, however, definitions of evaluation vary. Blaine Worthen and James Sanders define evaluation as "the formal determination of the quality, effectiveness, or value of a program, product, project, process, objective, or curriculum." Evaluation includes inquiry and judgment methods: "(1) determining standards for judging quality and deciding whether those standards should be relative or absolute, (2) collecting relevant information, and (3) applying the standards to determine quality."[34]

Abbie Brown and Timothy Green define evaluation as the process of judging, based on gathered data, the success level of an individual's learning, or a product's effectiveness.[35] According to Norbert Seel and Sanne Dijkstra, evaluation furnishes data that enable us to compare worth or value of two or more programs. It provides a basis or bases for selecting programs or determining whether they should be continued.[36]

Daniel Stufflebeam has defined evaluation as "the process of delineating, obtaining, and providing useful information for judging decision alternatives."[37] Collin Marsh and George Willis indicate that evaluation permeates all human activity. It deals with questions

such as: Is something worth doing? How well is it being done? Do I like doing it? Should I spend my time doing something else?[38]

Many view evaluation as critical inquiry, studying phenomena in order to make informed judgments. Kenneth Sirotnik and Jeannie Oakes expand on this concept of evaluation. They argue that we should inquire into the assumptions underlying the values that we hold, the positions that we advocate, and the actions that we undertake.[39] Most evaluators maintain that although the presence and importance of values cannot be ignored, they can be considered only within a particular context. We judge whether a program reflects its values and if those in charge of a curriculum have made their values explicit. Then we evaluate whether these goals have been attained. Sirotnik and Oakes advocate a type of critical inquiry that some have called *hermeneutics*. The dictionary defines hermeneutics as "the study of the methodological principles of interpretation."[40]

In taking a hermeneutic approach to evaluating curricula and their effects, an evaluator raises "deep" questions as to the educational program's value, worth, and merit. Certainly, we pose obvious questions as to what students learn. However, we also recognize that what students have learned is decided by people both inside and outside the immediate community. We judge the value of the opinions of those who decide what students are to learn and who determines levels of success. Evaluators who take a hermeneutic approach consider how well the educational program fits into the current climate.[41]

Measurement versus Evaluation

Sometimes educators confuse measurement with evaluation. Fred Kerlinger defined *measurement* as assigning numerals to objects or events according to rules.[42] *Evaluation* assigns value and meaning to measurement. For example, an evaluator might decide that a score of 70 percent correct answers means "passing" or "successful performance."

Measurement describes a situation or behavior in numerical terms. We make observations and then assign numbers to aspects of the observed phenomena.[43] For instance, a gym teacher can note the number of pushups a student does, or a reading teacher can record the number of pages per hour a student reads.

Measurement enables educators to record students' degrees of competency. However, educators must do something with the gathered data. They must decide how many pushups are enough to be good, and the extent to which reading *speed* equates to reading *ability*. They must decide whether a student who spells 18 of 20 words correctly should get an A, an A–, or some other grade. Measurement always precedes evaluation. The value judgments made in evaluation are always influenced by the educators' understandings of a program's—and education's—purposes.

APPROACHES TO EVALUATION

Evaluation is not content specific. The same procedures can be used to evaluate the effectiveness of any curriculum. Essentially, evaluation consists of gathering data and relating them to goals. In determining the value of a curriculum plan, educators must ask whether the expected results are worth the likely cost of delivering them.[44]

Scientific, Modernist Approach to Evaluation

How people generate questions and process data is influenced by their philosophy and psychology. Their philosophy and psychology are shaped by whether they consider themselves within the modernist or the postmodernist camp. Those who take a behavioristic, prescriptive, or sequenced approach to evaluation can be grouped in the modernist camp. They believe strongly in cause-and-effect precision in explaining the physics of the world and the exactitude of their actions in various endeavors, in our case the development, implementation, and evaluation of curricula and instructional strategies. These modernists approach evaluation specifying specific behaviors or content learned as a result of curriculum and instruction. They prefer clearly stated objectives

and precise indicators of whether their students have achieved the program's intended outcomes. They favor utilizing standardized tests to measure the attainment of learning objectives.

Humanistic, Postmodernist Approach to Evaluation

Educators who take a humanistic approach are more interested in whether the planned situations have enabled students to improve their self-concepts. They may not pay as much attention to students' specific achievements as indicated by objective tests.

Humanists, postmodernists, shun the thrust of modernity's search for truth and certainty.[45] They realize that evaluation cannot provide educators with precise results of students' learnings after experiencing various curricula and teaching strategies. Postmodernists shun employing scientific and precise measures of evaluation. They engage in "the art of interpretation."[46] They denote that their methods involve "intersubjective communication and answerability."[47] These evaluators employ various forms in interpretive inquiry. They rely less on statistical methodologies, preferring methods such as aesthetics, ethnography, autobiography, phenomenology, critical literacy, and various forms of heuristics.[48]

Slattery has critiqued modernists as searching for fundamental truths that can explain and quantify individuals, their unique experiences, even the workings of the cosmos.[49] Postmodernists engage in hermeneutic inquiry and evaluation, which reveal that the nature of life and the foci of our inquiries produce not certainty, but ambiguity, uncertainty, and risk.[50]

Doll argues that postmodernism requests that our educational actions, including evaluation, embrace a new educational posture. The only reality we have is the evolving present, which we all experience.[51] And, if we are attentive, we recognize that the here and now, our lived experience, is in disarray. Doll invites us to embrace this new approach to education, to curriculum building and curricular evaluation. However, he recommends patience in transitioning to postmodernism.[52]

Doll does suggest some stances we can take to commence our embrace of the postmodernist orientation. These perspectives impact not just curriculum evaluation, but all actions requisite for generating dynamic curricula and energetic instruction: celebrating doubt as we engage in curricular actions; stressing collaborative interactions with principal curriculum players; and critiquing our endeavors as we proceed. Essentially, Doll suggests that all participants in curricular engagements embrace the notion of dynamic interacting communities.[53]

Celebrating doubt directly challenges the modernist posture of elevating uncertainty. A modernist believes that a test score denotes a certain mastery or understanding of some subject matter. Further, he or she is convinced that students' high marks on a test indicate teacher effectiveness. A postmodernist realizes that test results and the effectiveness of a particular curriculum and various pedagogies are always open to diverse interpretations. One continuously engages in self-criticism and doubt. Pedagogical mastery and evaluative precision are illusions, essentially unattainable goals, such as arriving at an earthly horizon. Certainty eludes us in our every act. This also is true regarding students' engagements with their learning.[54]

Our interpretation of Doll's second stance is that the key players should stress collaborative interactions. Teachers do not, or should not, present monologues that their colleagues or students must accept. Educational activity and the myriad resultant learnings and dispositions of teachers and students result from interactions with others within the school and classroom cultures. The specifics of curricula and the means of evaluating them emerge from the dynamics among students and teachers. Novelty and surprise are embraced, and playfulness in the class and school communities is encouraged. Even evaluation can involve dialogue among students, rather than a monitored task performed alone and in silence.

We have modified Doll's third recommended stance: reinterpreting the practical to having all educational players, teachers and students, engage in continuous critiques of their endeavors as they plan and carry out evaluation. Focus and reflect on activities engaged. Teachers and students, do not have your actions hindered by theoretical constructs. Rather, study in depth what

occurs within the classroom, become participant observers of "educational theater" in the classroom and the local community. Incorporate your observations into a "playbook" for creating and evaluating the educational experiences of students as well as yourself and your colleagues.[55] If educators study in depth what occurs within the classroom, actually become participant observers of themselves and their students, they begin to recognize that "lived experiences" in the classroom and the local community actually can become the "playbook" for creating, enacting, and evaluating both the educational experiences of students and the effectiveness of the teacher. Also, if teachers bring into this stance the centrality of the dialogic process, they will enhance the collaborative nature of everyone's educational experience.

In Doll's final stance, he is urging all participants in "curricular theater" to engage in dynamic interacting communities. He refers to this theater as an ecological framework. Modernism celebrates individualism. It embraces a separation of ourselves from the environments we inhabit. Postmodernism honors our communion with others, persons, fauna, and flora. It accepts the notion that our realities are not static; they are always materializing. We as individuals are also evolving. Evaluation methods cannot stop this evolution. Test scores attained at one point in time cannot be accepted with certainty at another point in time. Individuals, teachers and students, exist within complexity and chaos. We need to realize that in the world community, we should not be in competition. We should recognize and embrace our communion. Teachers and students are mutually interdependent. Citizens of the United States are members of the world community that must champion "cooperative communalism."[56]

In general, evaluation enables educators to (1) decide whether to maintain, revise, or replace the existing curriculum; (2) assess individuals (primarily teachers and students) in terms of instruction and learning; and (3) decide whether the existing managerial organization of the school and its program should be maintained or reformed. Also, part of evaluation focuses on the school environment and the community environment within which the school exists.

Richard L. Curwin posits another reason for engaging in evaluation. He cites the value of employing evaluation as a means of motivating students to increase their learning. He indicates many educators believe that successfully achieving some learning goal leads to motivation. He purports that is backward. "Motivation or effort leads to success, not the other way around."[57] For motivation to stir learning, it has to present some type of challenge to the student. Previously, we compared a video arcade to a classroom. In the arcade, students just learning the game often fail to attain high scores or success. But the game has challenges that do not discourage the gamers; rather, it motivates them to attempt again, to increase their efforts to be successful. Learning should be playful and challenging; it should furnish data to students to know how they are doing. It should present opportunities that challenge students that stimulate in them an increased desire to learn. Curwin defines "the desire to learn" as educational motivation.[58] The most important aspect to curriculum evaluation is not to sort students or teachers, but to foster in students a thrill of and a perseverance in their learning and a record of their educational journeys.

Of course, evaluation occurs at different levels. But we argue that the process regardless of levels should serve the primary purpose, to let students and teachers and even the community gain data that excite the mind, motivate learning, and stimulate a love of learning. Also, evaluation should not discourage, but encourage students to play the learning game, relish what they know and be thrilled to engage in knowledge strategies that delve into what they realize they do not know. Evaluation should whet the appetites of the mind and the spirit.

At the broadest levels, evaluation focuses on an entire school district, state educational system, or even national system (e.g., with regard to No Child Left Behind legislation). Narrower evaluation focuses on particular institutions, either individually (e.g., a particular high school) or a group (e.g., all the high schools within a particular district).

At the most specific level, evaluation attends to a particular program for a particular course at a particular grade level. What is valued at a broader level should also be valued at a narrower level. It makes no sense to indicate that U.S. schools will be judged according to particular criteria if schools at the local level reject or cannot feasibly apply those criteria.[59] In 2002,

No Child Left Behind mandated that all students, even those with learning disabilities or limited English-language competency, be held to the same standards as the regular school population. They had to pass tests in reading and mathematics. Educators and others noted then and continued to protest that it was unrealistic to expect students with limited or no ability to speak English to pass a test written in English. It was also not realistic to assume that children with limited intellectual capacities could achieve at levels comparable to average children.

The U.S. Department of Education began to listen. In 2004, the department altered the rule, enabling first-year immigrants to opt out of taking the reading test. However, they still had to take the state's mathematics test. Their reasoning appears based on the fact that many students, especially Asian students, with limited English skills still do quite well in mathematics. In 2007, the Department of Education admitted that there would always exist a small number of students whose abilities are such that it is not possible to assess them meaningfully. School districts are allowed in certain cases to use alternative standards of assessment or developmentally appropriate versions of the state assessment.[60]

Scientific, Modernist Approach versus Humanistic, Postmodernist Approach

Lee Cronbach places scientific, modernist and humanistic, postmodernist approaches at opposite ends of the evaluation continuum. Actually, Cronbach does not use the terms modernist and postmodernist; we have made this adjustment. And we are not sure that these two approaches are at opposite ends of an evaluation continuum. Rather, it appears that the scientific, modernists, rather than being in a versus category with humanistic, postmodernists, are morphing into a new 21st century way of contemplating life, education in our case.

However, we the authors believe that we certainly have not left the scientific, modernist posture, but that we are tweaking it in some cases. Those in the modernist camp do favor an experimental approach to evaluation. "(1) Two or more conditions are in place, at least one of them being the consequence of deliberative intervention. (2) Persons or institutions are assigned to conditions in a way that creates equivalent groups. (3) All participants are assessed on the same outcome measures."[61] They use data, frequently in the form of test scores, to compare students' achievement in different situations. Data are quantitative, so they can be analyzed statistically. Program decisions are based on the comparative information gathered.

Most scientific approaches to evaluation draw on methods used by physical scientists. Objective tests, a hallmark of traditional approaches, are still major vehicles by which educators gather data. Of course, with further research on evaluation, essay exams and other forms of gathering data are being employed within the scientific camp. Data tend to be quantitative, but this is changing. Often program decisions are based on the comparative information gathered, but evaluators are beginning to realize the shortcomings of just using data to compare students' achievement levels. This has been noticed previously.

Catherine Taylor and Susan Nolen mention that within the scientific camp, people make four assumptions that are, in reality, problematic: (1) students are randomly assigned to schools, teachers, and curricula; (2) instruction is identical for all students in the "treatment" condition; (3) some students will have positive learning experiences from the treatment, and other students will not; and (4) objective tests are accurate and impartial judges of students' learnings and skills.[62]

Taylor and Nolen note that educators cannot blindly accept these assumptions for the following reasons: (1) students are not randomly assigned to districts, schools, programs, or teachers; (2) rarely is instruction identical for all students, even in the same school or classroom; (3) treatments in classrooms do not remain constant; and (4) tests are not impartial.[63] These authors expand on why these assumptions must be challenged. The geography of school districts and the policies of school placements are not driven by a desire to create random groups of students. Schools serve most often the students within an attendance region. Teachers realize that they individualize their instructional strategies and educational activities, even when teaching

the same curriculum. A creative classroom has great diversity of the teacher's and students' actions. Also, effective teachers strive to be an educator of many "notes," not just a "Johnny-one-note." Teachers know that tests as they are designed address various students' academic strengths and even cultural backgrounds. Students who do well on multiple-choice tests are often highly skilled in memorization and recognition. Students have various learning styles, and tests usually do not stress several learning styles.[64]

It certainly appears that the high-stakes accountability environment in which we find ourselves does favor some version of the scientific, modernist approach to evaluation. The No Child Left Behind legislation seems to be forcing educators to hold supreme objective exams, and even subjective exams in some instances, to document that educational programs developed and delivered are attaining desired results. Gina Schuyler Ikemoto and Julie Marsh note that schools and educators are realizing that data-driven decision making (DDDM) is central to proving accountability and the meeting of standards. However, Ikemoto and Marsh caution that we must not assume that DDDM is a rather straightforward process. They point out, and support it with their research, that there is variety in the ways in which educators at school levels use and interpret data.[65]

Ikemoto and Marsh assert that DDDM in evaluation can be influenced by two conditions: the type of data gathered, and the approach or approaches to data analysis and decision making. In the DDDM process, educators can process a plethora of various types of data that can go from simple to complex. Simple data are less complicated and inclusive, usually focusing on only one specific aspect of a particular subject. Usually, evaluators dealing with a less complicated evaluative focus bring only one perspective to the analysis. Those dealing with complex data tend to view the evaluation situation as multidimensional. In such situations, evaluators draw on both quantitative and qualitative data. Here we see a blurring of scientific, modernist and humanistic, postmodernist approaches to evaluation. We submit that perhaps centering on these two camps of evaluation really does not serve us well. We should not worry about classifications of evaluation, but rather, we should focus on those strategies that enable us to gather evidence that answer the question: Is what we are doing in delivering this curriculum successful in attaining our goals?

These researchers note that the evaluative process, as mentioned previously, is also influenced by the type of decision making regarding the data gathered. They assert that the types of decision making also follow a continuum from simple to complex along several dimensions: "basis of interpretation (use of assumptions versus empirical evidence); reliance on knowledge (basic versus expert . . .); type of analysis (straightforward techniques, such as descriptive analyses, versus sophisticated analyses, such as value-added modeling); extent of participation (individual versus collection); and frequency (one time versus interactive)."[66] James Comer added: "All the money we spend on research, training, equipment, instructional programs, and the like will give us too small a return on our investment until we help the adults working together in a building learn to create a culture in which they can collaborate with each other in a way that will support the development of students."[67]

A major challenge in this century, relating to creating a culture in which individuals with different philosophies and orientations toward life and education can collaborate to support the total development of students, is that most educators really do not know whether they are modernists or postmodernists.[68] They cannot ascertain just how they view the world. Many who have heard about postmodern thinking experience difficulty in conceptualizing this orientation to varied realities. They cannot embrace using ambiguity and uncertainty to comprehend, much less evaluate, an ever chaotic and changing educational reality. This is not a critique of educators or the general public. We have lived in the modernist world since the Enlightenment. We have adopted Newtonian physics as our model. Now the world is being turned upside down. Our major premises are being disputed. Postmodernists proclaim imperfections are not failures but goals that serve to motivate innovative actions. Yet, these actions cannot guarantee better "futures." Doubt is always our companion, and that is how it should be.[69]

Rather than educators trying to classify in which approach to evaluation they are, it might better serve them to realize that they function in an evaluative culture that they must nurture. In order to be effective, educators must assess the effectiveness of the curriculum and its delivery. Evaluative data, whether gathered in a scientific or a humanistic frame, provide guidance for the continuation or the cessation of action regarding the curriculum. School cultures must foster not only creativity in creating curricula, but also creativity in evaluating the curricula and the instructional strategies embraced. Teachers must embrace the collaborative model of teaching. Teaching is not a solitary series of actions performed behind closed doors. We advise that schools foster a culture that enables the sharing of data, instructional ideas, and evaluative data so that school curricula are determined successful in stimulating total student growth.

Having said that, it might be more useful for educators to realize that they exist in an evaluative culture rather than to try to classify themselves as either scientific, modernist or humanistic, postmodernist. It behooves educators to realize that their approach to their school cultures is colored by whether they view data gathered on the effectiveness of curricula from an accountability culture or an organization-learning culture. If educators subscribe to the first view, they gather data to assert that the curricula offered raise test scores. Higher scores define curricular and instructional success. Those who embrace the organization-learning culture view test results not as an endpoint, but a way point, to indicate that the curriculum is contributing to the students' educational advancement.[70]

Educators who adhere to the accountability culture value a polishing of student understanding, efficiency of instruction and learning, and an immediate identification of learning. Those in the organizational-learning culture consider education as a dance, or a movement in motion between teacher and students. This posture celebrates adventure, "discovery, risk-taking, and long term development."[71]

Of course, we need not take sides. We can have allegiance both to accountability and to organization learning. However, as William Firestone and Raymond Gonzalez point out, districts tend to be drawn to one or the other philosophical orientation.[72] The camp to which people are drawn has intended and unintended consequences that influence how they view students and their learning, how they view themselves as educators, how they use data gathered, how they reflect on how time is processed within the evaluation process, and how teachers and administrators view their interactions in curricular activity, specifically evaluation.[73] As previously mentioned, a school culture that stresses an accountability culture primarily centers on test scores as the ultimate indicator of student learning. What do the students know? A school culture tending toward an organizational approach to evaluation is more interested in utilizing data that furnish information that enables an improvement in student learning.

In an accountability culture, teachers employ data to determine how well they are teaching and how well students are learning. Do data indicate that teachers are in compliance with district, state, or national edicts? Schools stressing organizational learning are more concerned with improving learning and curricular experiences. The stress is on way points, not endpoints. Rather than just reporting that data indicate that students have learned, a school with the organizational-learning culture wants to know not only if students are learning, but why they are learning, and if not learning, why not. In this latter camp, data are employed in diagnostic manners.

Educators stressing accountability consider the time frame to be essentially short term. Educators within the organizational camp realize that student success takes time. The accountability emphasis in evaluation favors a top-down organization. Data are directed to the central office or the office of evaluation or research, where they are processed. After analysis, information and guidance are issued down the chain of command. Organizational cultures are horizontal. Colleagues behave more like learning communities, mutually analyzing data and suggesting educational approaches or curricular content that might improve student learning.[74]

The organizational school culture tends toward utilizing humanistic, postmodernist approaches to evaluation. Students and teachers are not test-taking or test-giving machines. Students are not one-dimensional individuals. Educational colleagues likewise are not

one-dimensional. Although important, tests and their scores do not reveal the entire story. And, where tests are used to compare and rank students, the tests might not provide any information of value. It appears that people are increasingly interested in more humanistic approaches. People are realizing that nontraditional evaluation procedures may furnish more complete pictures of curricula. The humanistic approach, although not completely rejecting objective tests, stresses that educators can gather more useful data employing more naturalistic approaches such as case studies and participant observations. Educators of this stripe prefer to study programs already in place rather than programs imposed by groups outside of the school district.

Humanistic evaluators primarily analyze qualitative data, such as impressions of what they observe. They describe actual incidents. They gain data by interviews and discussions with participants, students and teachers included. Analysis seeks to uncover patterns among many observations.

Those advocating the humanistic, postmodernist approach to evaluation argue that this approach is necessary at a time of multiple voices and multiple realities. We must make judgments about the complexities we find within the educational system and within the general society. And these judgments must be tentative; we cannot arrive at judgments and conclusions with abstract and generalized certainties, as advocates of the scientific approaches would have us believe.[75]

Although various models are employed in the traditional quantitative camp, most seem not to have particular names. Such is not the case with approaches to qualitative evaluation and research. We discuss five major humanistic approaches that have been identified: interpretive, artistic, systematic, theory driven, and critical-emancipatory.[76] While we have clustered these approaches within the postmodern realm, we are aware that advocates of the approaches might disagree. Such is the case because postmodernism is in a state of flux; it is continually emerging; it is constantly engaged in self-reflection, self-analysis, constantly attempting to engage uncertainty, chaos, and complexity.

In the *interpretive approach,* the evaluator considers the educational scene and interprets the meaning and significance of peoples' actions. Attention to social context is essential. The evaluators are people directly involved with the curriculum, especially teachers and students.

In the *artistic approach,* the evaluator engages in aesthetic inquiry, observing classes and other enactments of curricula and then publicly announcing what is good and bad about the curriculum. This approach relies on individual intuition honed by experience.[77] The evaluator focuses on the quality of the relationships between teacher and students. The key advocate of this approach is Elliot Eisner, a former professor emeritus of art and curriculum at Stanford University.

Among humanistic approaches to evaluation, the *systematic approach* is most familiar. Evaluators try to be as objective as possible in their descriptions, employ logical analysis and base their judgments on fact. However, they do not rely primarily on statistical techniques, the hallmark of the scientific approach.

Many evaluators take a *theory-driven* approach. These calculators apply philosophical, political, or social theories when judging the quality of curricula.

Critical-emancipatory evaluators tend to be the most radical. They judge a curriculum's quality and effectiveness according to how well the curriculum counters social forces that impede individual development and fulfillment. These evaluators draw heavily on Jurgen Habermas's work on the construction of knowledge and meaning. They also draw on critical theory, especially Marxist theory.[78]

Educators need not be tied to any one of these five major approaches. Indeed, there are several other ways to identify the approaches to evaluation.

Utilitarian versus Intuitionist Approach

Evaluation can be classified as either utilitarian or intuitionist. The utilitarian approach is closely linked to the scientific, modernist approach, whereas the intuitionist approach is tied to the humanistic, postmodernist approach.

Utilitarian evaluation operates according to the premise that the greatest good is that which benefits the greatest number of individuals.[79] Utilitarian evaluators look at large groups, such as an entire school or school district. Attention is on total group performances. Programs are judged by how they affect the school's overall student population. Programs that allow the most students to attain the objectives are judged worthy of continuation. *Intuitionist evaluators* gather data to judge the program's impact on individuals or small groups. There is no one criterion regarding worth. Numerous criteria are employed to assess a program's worth. Program participants, not outside evaluators, consider the program's quality. Everyone affected by the program can make judgments about it.[80]

Intrinsic versus Payoff Approach

In addition to viewing evaluation in terms of scientific, modernist versus humanistic, postmodernist or utilitarian versus intuitionist, we can view it in terms of what Michael Scriven has called intrinsic versus payoff.

Intrinsic evaluators study the curriculum plan separately. Their evaluation criteria are not usually operationally defined. Instead, the evaluators are merely trying to answer the question, "How good is the curriculum?"[81] Intrinsic evaluators study the particular content included, the way it is sequenced, its accuracy, the types of experiences suggested for dealing with the content, and the types of materials to be employed. They assume that if a curriculum plan has accurate content and a firm basis for its particular organization, it will effectively stimulate student learning. All evaluators must engage in intrinsic evaluation—that is, they must determine if the curriculum has value. Evaluators must consider not only how well a course or curriculum achieves its goals and objectives, but whether those goals and objectives are worthwhile.

Once a curriculum's basic worth has been assessed, evaluators must examine the effects of the delivered curriculum. This is *payoff evaluation*. Often, the outcomes are operationally defined. Evaluators can consider the curriculum's effects on students, teachers, parents, and, perhaps, administrators. This evaluation approach may involve judgments regarding the differences between pre- and posttests and between experimental-group and control-group tests on one or more criteria parameters. Payoff evaluation receives the most attention from educators because it indicates curriculum's effects on learners in terms of stated objectives.

Supporters of the intrinsic approach agree that important values cannot be assessed via the payoff approach because of deficiencies in present test instruments and scoring procedures. Also, the results reported in payoff evaluation studies are usually short-term results of a curriculum. Little attention is given to a program's long-term outcomes. If educators want to have an idea of a curriculum's relevance and perhaps elegance, they would do better to look at the curriculum's materials directly rather than at students' test scores.

Formative and Summative Evaluation

Another way to view evaluation is in terms of formative and summative evaluation. *Formative evaluation* encompasses activities undertaken to improve an intended program—that is, optimize student learning. Formative evaluation (sometimes called *rapid-prototype evaluation* by instruction designers) is carried out during program development and implementation.[82] In the curriculum-development phase, formative evaluation furnishes evidence that directs decisions about how to revise a program while it is being developed. Formative evaluators look at specific subunits of the curriculum being developed and test them in brief trial situations. They gather data, often in classrooms, that inform their decisions about how to modify these program elements before they are fully implemented. During a curriculum's developmental and early piloting stages, formative evaluation provides frequent, detailed, specific information. Formative evaluation takes place at a number of specific points in the curriculum-development process. It is essential, especially during the initial stages of the development process.[83] Formative evaluation allows educators to modify, reject, or accept the program as it is evolving.

How educators conduct formative evaluation varies widely. If they are evaluating only one unit plan, their manner of evaluation may be very informal, perhaps involving only the people teaching the unit. However, if they are engaged in creating a new program for an entire school district, formative evaluation may be more formal and systematic.

Formative evaluation also occurs during the teaching of a new or existing curriculum, focusing on teachers as well as students. Teachers can use formative evaluation to judge the effectiveness of their pedagogical approaches. Teachers must realize that formative evaluation is not a sometime activity. It is a grand composite of ways to gather and utilize data in order to make those instructional adjustments necessary for optimal student learning. Such evaluation furnishes feedback to the teacher as to how a lesson is going and how it might be fine-tuned.

For teachers to fine-tune their pedagogical strategies, they need to utilize formative evaluation to assess the levels of students' learnings and understandings. Brent Duckor denotes that teachers need to realize that formative assessment is not just a cluster of teacher-made or even standardized tests. It is much more than a checklist of student qualities. It is more than a file of collected student activities. He points out that it involves a series of teacher and student moves that define a continuous relationship between students and teacher.[84]

He outlines several essential moves that teachers can take to enhance engaging in formative assessment. Essentially, these steps involve a questioning strategy. The first step presented is to *prime the students* that you, the teacher, will be raising questions that will engage students in deep reflection. One-word answers will be insufficient. Also, students will be expected to challenge fellow students as to why they answered as they did. Step two is to *ask effective questions*. This means that questions should do more than demand knowledge responses. Questions need to address all the levels of Bloom's Taxonomy, Cognitive Level and Krathwohl's Taxonomy, Affective Level. Allow students *time to ponder the question,* time to generate an in-depth response. This is Duckor's third move.

The fourth move in this questioning formative evaluation strategy is to not let students "off the hook" with a quick acceptance of the answer. Formative questioning is not to mine certainty; it is to *probe for a rich response,* exposing deeper apprehension. Here both teacher and class members are deeply engaged in formative evaluation, assessing what is known and what is now recognized as not known. The fifth step in formative question evaluation is to *distribute the questions among all class members.* Through such questioning by both teacher and students, a record of responses is recorded. Later analyses can reveal how understandings over time have advanced. Answers can be categorized as to their value in advancing one's comprehension and even creativity.[85]

Frederick Erickson notes that for formative evaluation to really occur, teachers must know how to interpret the data gathered. Lacking interpretative understanding prevents the teacher from making instructional adjustments. Erickson asserts that often teachers are not skilled in analyzing and comprehending data. Thus, no formative evaluation occurs. Even if teachers do know how to apprehend the data, they often lack the time for analysis. It seems that perhaps a majority of teachers feel the need to "cover the book" in a certain time period. It takes time to self-critique and make pedagogical adjustments. Teachers often report that they do not have time to reteach a lesson. That objective test must be administered on time. There is so much content to teach; so much content is on the test.[86]

Erickson argues that we cannot just mandate that teachers employ formative evaluation; we must schedule time for them, working alone and with colleagues, to raise questions about what the data are telling them. He points out that teachers must really possess pedagogical content knowledge at a deep level. Skilled teaching is complicated, and often it represents improvisational theater in which teachers have to pick up on classroom dialogue from the questions and statements of students. Pedagogical content knowledge, we declare, is not solely in the domain of educational methods or instructional strategies. Pedagogical content knowledge is essentially drawn from procedural knowledge associated with the declarative knowledge of a discipline of study. Essentially, pedagogical content knowledge draws and adapts its techniques from the ways

in which scholars actually advance their understandings within their specific fields. Biologists use specific methods to advance biology. These methods differ from how historians advance their understandings of some historical period or event. Mathematicians engage in processes of solving problems unique to their fields of expertise. For example, a biologist who seeks to prove the validity of some experiment does not argue the case in point as would a historian or a mathematician. A biological investigation differs greatly from a historical inquiry. If we wish students to learn biology, history, or mathematics, our instructional methods must mimic the ways that experts in these fields also go about their learning.

Of course, experts in various disciplined fields often engage in interdisciplinary activities. Thus the biologists often utilize mathematics in experiments. Understanding this and the range of fields of study makes it even more challenging for teachers to teach—or, more accurately, to get students to learn these subjects.

Formative evaluation also refers to procedures employed by students to assess their learning tactics as well as their levels of knowledge.[87] Students must know what they know and how well they employ particular learning strategies. The level of student involvement in formative assessment depends, of course, on their maturity levels. However, even students in primary school have some idea as to whether they understand something. They certainly need the teacher's guidance to determine ways to approach learning. We want our students to become independent learners. As students gain more expertise in learning and greater knowledge, they can assume more management and refining of their learning adjustments. As W. James Popham indicates, teachers take on a more supporting role in suggesting ways to learn more effectively.[88]

Today, as more and more schools are establishing computer-based learning environments, they are actually employing formative evaluation or assessment. As Allan Collins and Richard Halverson indicate, these computer programs embedded formative assessment into the actual lessons. As students proceed through the computer curriculum, the computer furnishes feedback indicating either progress or where an error has occurred. If an error is indicated, the computer program maps out a strategy or strategies to correct the error or arrive at a correct answer. Essentially, the computer can be assessing the cumulative results of particular learnings of knowledge and strategies. In interacting with the computer program in this way, students realize that making a mistake actually provides an opportunity for immediate learning. With such feedback and really no grade on the line, students avoid taking misinformation or misunderstandings into their further learning.

We point out here that the computer is not replacing the teacher as instructor or evaluator. It is merely enabling the teacher at times to reach more students.[89] As classrooms become more like "learning laboratories," teachers and students become highly involved in the learning and evaluation processes, more engaged in dynamic interactions with each other and evolving "technological assistants." Technology will, we believe, actually humanize the teaching-learning process. Also, technology means that teachers, students, and even expert evaluators need not always be in the same physical spaces.

Of course, one need not abandon relatively mundane means of gathering formative evaluative data. Mike Schmoker discusses the power of a quarterly curriculum review in which supervisors—and, we would add, teachers—meet to discuss "how things are going." Discussions could focus on periodic formative assessment, team lesson logs or learning logs, and particularly samples of students' work. All participants can get a "feel" for how things are going, the effectiveness of particular curricular units, and the power of certain pedagogies and student class organizations.[90]

Taylor and Nolen list various assessment tools that are not high tech: anecdotal records, checklists, rating scales, conferences, journals, even homework.[91] They also note that teachers can engage in formative assessment just by walking around the classroom and observing and listening to students. Much data can be obtained when teachers listen in on brainstorming. Even wish lists regarding topics to be covered can be employed. Having students enumerate what they like to do when away from school can furnish much evaluative data useful in planning future lessons.

Summative evaluation is aimed at assessing the overall quality of a produced and then taught curriculum. As Wilhelmina Savenye notes, data are gathered to ascertain the new program's worth and effectiveness.[92] If formative evaluation has been implemented carefully, summative evaluation should indicate that the program has enabled students to attain the curriculum goals. Such summative evaluation informs educators that students have met the school's or state's educational standards. It also indicates that teachers have met the minimum accountability standards.

Overall, summative evaluation poses the question, has the curriculum worked? As its name implies, summative evaluation gathers evidence about the summed effects of a particular curriculum's components or units. We issue a note of caution regarding the question of whether the curriculum has worked. Ideally, we will find that it has worked, but only in "small letters." There are still multiple levels to the statement "worked." Here we have placed quotation marks around "worked" to emphasize, as does Doll, that the data at hand, while useful, have to be considered always partial.[93] Whether you interpret this caveat as postmodern or modern, we educators must realize that we should never be satisfied with our answers to questions, our documentation of our effectiveness, or our reporting of our students' learnings and mastery with absolute certainty. The results of evaluation, especially summative evaluation, are not endpoints, but way points. Our education, our actions, our evaluative assessments are endeavors situated, as Doll asserts, in realities continuously emerging in divergent directions.[94]

Brown and Green discuss an approach to summative evaluation that D. L. Kirkpatrick developed in the mid-1990s. Although Brown and Green are discussing summative evaluation in terms of instructional design, Kirkpatrick's approach can be applied to curriculum evaluation. Kirkpatrick delineates four levels of summative evaluation: (1) reactions, (2) learning, (3) transfer, and (4) results.[95]

Level 1, reactions, focuses on gathering data about how students reacted to the new program. The data indicate not only the amount of new knowledge acquired, but also whether what was provided to students was relevant to them. Did the new curriculum and attendant experiences meet students' social, emotional, and intellectual needs? Did the students react in anticipated ways? At level 1, evaluators might interview students or have them respond to attitude surveys (rather than tests).

At level 2, evaluators gather data on whether students have gained new knowledge, skills, and techniques implicit in the new program's goals and objectives. To collect such data, evaluators usually administer a series of pretests and posttests at various junctures of the implemented curriculum.

At level 3, evaluators pose questions about whether the individuals who experienced the new program can effectively employ newly acquired skills and knowledge and whether their attitudes have changed for the better. Using various types of tests, evaluators determine if students show evidence in everyday life, job situations, or further schooling that they are applying their new knowledge, skills, and attitudes.[96]

Level 4, results, is a major challenge for evaluators. The results of a newly developed curriculum may not be evident immediately, if ever. Some schools assess results partly through exit interviews of students, which indicate how the new curriculum has changed their knowledge, skill, or attitudes. Evaluation at this final level might also be conducted via focus-group activities. Surveys given to graduates of new curricula can also furnish summative data.[97]

The results of summative evaluation present not just a major challenge for evaluators, but a multitude of challenges for all concerned with the total educational "theater." Many educators and the general public are not even aware of these challenges, largely because most of us rarely question our conceptions of world realities. We educators take for granted that we truly comprehend the essential natures of teaching and learning. Accepting that, we neglect to reflect deeply on just what they are. Can we really know their nature?

In summative evaluation, it is assumed, usually without challenge, that teaching is an activity that can be accomplished in a specific time frame. Likewise, learning also exists in time. We can finish teaching a unit. Students can finish learning a particular lesson. We can, in our

evaluative roles, create summative tests given at a specific time that can accurately document a level of understanding or accomplishment. And we can make, from analyzing the test data or score, that "'what is learned' by students is . . . an entity that comes to exist after instruction has taken place, and thus, can be measured as a whole thing of the past. This ontological presupposition is the foundation for the entire enterprise of summative evaluation."[98]

Some, if not most, advocates of summative evaluation also assume that formal psychometric procedures are essentially the best way to gather reliable and valid documentation of student learning. Essentially, we cannot trust that teachers in the classroom using observational and other formative measures will furnish us with results that can inform us as to what educators are doing.[99]

Teaching is never completed, nor is it performed only by the teacher. Also, teaching at times occurs outside the classroom or school environment. Time is fluid with teaching. Likewise, complete learning, sometimes called *mastery learning,* in reality is never attained. Learning is ongoing, never ceasing to enrich understanding. Certainly learning exists, just as a horizon on an ocean exists. However, most of us know that we can only advance toward the horizon; it can never be reached.

And if we could somehow magically reach that horizon, our voyage would be over. Likewise, if we could really attain mastery, then our education, our journey of learning, would cease. Learning is the result of ongoing interactions with numerous peoples in a multitude of environments. Erickson notes that learning is "the process of acquisition itself, as continual change within an ongoing course of activity."[100] In this view, learning to know and comprehend the content of the curriculum represents beginnings and way points, not endpoints that can be precisely noted and statistically analyzed.

In summative evaluation, attention is on demonstrated results—on acceptance of an audit culture.[101] Summative evaluation essentially ignores the subjective aspects of learning, the emotional valences students possess. It is difficult to have a summative test for thrills or infatuation.

As Taubman notes, the learning sciences have and continue to strive for objectively measuring learning, writ large. Essentially, learning sciences are not concerned with intrinsic evaluation of the worth of curricular content or curricular experiences. Learning sciences seem enamored only of getting students to learn and urging teachers to teach.[102] After all, we hear often that if we just had good teachers and good schools, students would learn and would be prepared to compete in the world marketplace. Few question that if we just had a quality curriculum, if we just had highly emotional experiences, students would truly be changed. It is hard to measure that summatively.

The preceding discussion is not to discount summative evaluation procedures, but rather to enlighten us that even if we could create the perfect summative test with reliability and validity, we will still have only an incomplete portrait of what students have learned and teachers have taught. Much of learning and teaching will never be known, and the mysteries around these human interactions are to be celebrated. All evaluations, both formative and summative, are to be enacted with an awareness of their pluses and minuses. Education is not engineering; it is far more complex. We know when a building is complete. In education, we never know what it means for a person to be complete. Humans never attain completeness.

We hope that you, the reader, realize that the next section, evaluation models, is to be processed essentially as descriptions of evaluative procedures. It is to be hoped that the models contain within them explanatory elements.[103] Also, keep in mind that although the models may present a clean procedural pathway to gather data and make decisions, in actuality, the models can and do get messy when actually employed.

EVALUATION MODELS

Previously it was noted that evaluation was not content specific, and the same or similar strategies can be employed to evaluate the effectiveness of any curriculum. However, the various approaches (scientific, modernist and humanistic, postmodernist) can and do influence the

assumptions evaluators consider when analyzing particular curricula and varied pedagogical strategies. These assumptions are embedded within philosophical, educational, social, and world views. Thus, while strategies utilized in assessment have similarities, there are distinct evaluation models under the scientific, modernist organizer and the humanistic, postmodernist framework.

Scientific Models, Modernist Models

The first large-scale formal evaluation in the United States was reported in Joseph Rice's 1897–1898 comparative study of the spelling performance of more than 30,000 students in an urban school system. Soon after, Robert Thorndike was instrumental in getting educators to measure human change.[104] Finally, the Eight-Year Study (1933–1941) was a turning point in educational evaluation, ushering in the modern era of program evaluation.[105] The Eight-Year Study's evaluation plan was organized in seven sequential steps: focusing on the program's goals and objectives, classifying objectives, defining objectives in behavioral terms, finding situations in which achievement can be demonstrated, developing or selecting measurement techniques, collecting student performance data, and comparing data against objectives.

STAKE'S CONGRUENCE-CONTINGENCY MODEL. Robert Stake distinguishes between formal and informal evaluation procedures. Although recognizing that educational evaluation continues to depend on casual observation, implicit goals, intuitive norms, and subjective judgment, he notes that educators should strive to establish formal evaluation procedures. Formal procedures are objective and supply data that enable descriptions and judgments regarding the program being evaluated.

Evaluators seem to be increasing their emphasis on providing full objective descriptions and on collecting and reporting hard data. Stake asks that evaluators collect and process more extensive types of data, consider the dynamics among people involved in the curriculum process, assess the roles various people play, allow those people greater participation in judging programs, and take positions regarding a program's worth.

Stake delineates three data categories: antecedents, transactions, and results. Applying this organization to modern-day evaluating processes yields three new categories: prerequisites, curriculum, and results. Prerequisites refer to any condition that exists prior to teaching and learning that may influence outcomes. *Prerequisites* include the status or characteristics of students prior to their lessons: their aptitudes, previous achievement scores, psychological profile scores, grades, discipline, and attendance. Prerequisites also include teacher characteristics such as years of experience, type of education, and teacher-behavior ratings.

Curriculum in the model refers to the planned or potentially considered interactions among students and teachers, students and students, and students and resource people. Curriculum also addresses students' potential interactions with curriculum materials and classroom environments. At this stage, educators attend to how the planned curriculum is affected by time allocation, space arrangements, and communication flow. Attention essentially is directed at the teaching process. In the curriculum planning stage, educators contemplate how the engagements considered actually play out when the curriculum is applied and evaluated.

Results are the program's anticipated and then acquired outcomes, including student achievement and, sometimes, attitudes and motor skills; impact on teachers' perceptions of their competence; and influence on administrators' actions. Evaluators must also consider long-range results and other outcomes not evident when a program concludes. According to Stake, educational outcomes are immediate and long-range, cognitive and affective, personal and communitywide. Stake's evaluation model encompasses curriculum design, development, and implementation. Data elucidate disparities between what was planned and what has actually occurred.

Figure 9.2 shows the deliberate connection of the prerequisites, curricula, and results in the planning stage. The evaluator looks for empirical information in the implemented curriculum. Do the data reveal that transactions are supported empirically in the implemented curriculum?

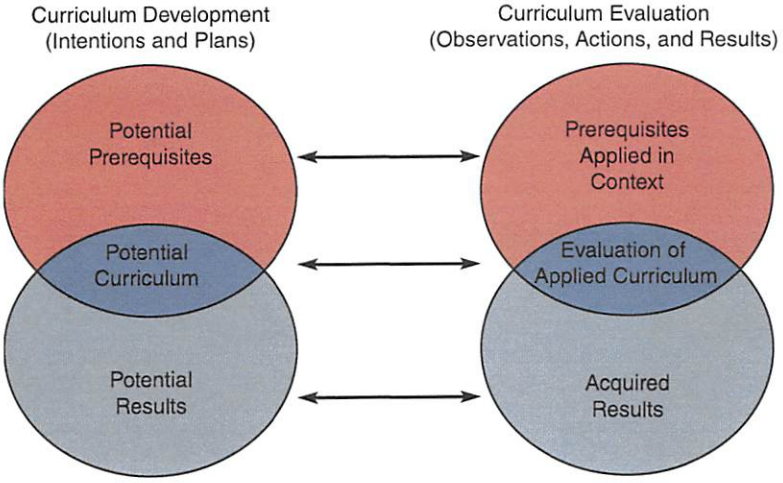

FIGURE 9.2 Consequence-Contingency Model

Source: Based on Robert E. Stake, "The Countenance of Educational Evolution," *Teachers College Record* (1967), p. 7.

Do data make the case that the results attained are really the consequence of the procedures employed during instruction? Effective evaluation links prerequisites, curriculum, and results in both the planning and evaluation stages.

Stake's model also depicts the relationships between what is planned and what is enacted and then evaluated. For complete congruence between plans and results, all observed prerequisites, curricula, and results must be the same as the intended ones. Although Stake's model is very useful, complete congruence is impossible. There is no exact correspondence between some action and student learning. Outside of school, students encounter material that affects their thinking about a particular lesson. Such an unintended transaction can result in learning noted as an attained outcome.[106]

STUFFLEBEAM'S MODEL: CONTEXT, INPUT, PROCESS, AND PRODUCT. Daniel Stufflebeam provides a comprehensive evaluation model that is an important contribution to a decision-management approach. According to Stufflebeam, information is provided to management for decision making. Evaluation must include the following: *delineating* what information must be collected, *obtaining* the information, and *providing* the information to interested parties. Stufflebeam delineates four types of evaluation: context, input, process, and product.[107]

Context evaluation involves studying the program's environment. Its purpose is to define the relevant environment, portray the desired and actual conditions pertaining to that environment, focus on unmet needs and missed opportunities, and diagnose the reason for unmet needs. Context evaluation is not a one-time activity; it continues to furnish information on the total system's operations and accomplishments (see Curriculum Tips 9.1).

Input evaluation provides information regarding resource use. It focuses on feasibility. Evaluators assess the school's ability to carry out evaluation. They consider the suggested strategies for achieving program goals, and they identify the means by which a selected strategy will be implemented. They might consider alternative designs that lead to the objectives while requiring fewer resources, less time, and less money.

Evaluators assess specific aspects or components of the curriculum plan. Input evaluation addresses these questions: Are the objectives stated appropriately? Are they congruent with the school's goals? Is the content congruent with the program's goals and objectives? Are the instructional strategies appropriate? Do other strategies exist that could achieve the objectives? What is the basis for believing that these contents and instructional strategies will result in attainment of the objectives?

CURRICULUM TIPS 9.1 Assessing the Curriculum Context

Most curricular actions occur within a socialized context, and most of their delivery or enactment processes take place within a socialized context. Those in charge of the overall program must evaluate the process by which they create and deliver curriculum. The following tips can assist in assessing the context of curricular action:

1. Determine the values, goals, and beliefs that drive the curriculum.
2. Obtain a reading of the community, noting the key players.
3. Determine the history of past curricular activity in the school district.
4. Get some indication of the physical facilities available and necessary for enactment of the curriculum.
5. Judge the pressures for and against actions generated from within and from without the community and school district.
6. Determine the budget needed and the budget allocated.
7. Determine what performance outcomes are important for the school and community.
8. Get a fix on the perceptions, expectations, and judgments of teachers and administrators, what they expect out of the evaluation, and how they intend to use it.

Source: Personal paper, F. P. Hunkins, 2005.

Process evaluation addresses implementation decisions that control and manage the program. It is used to determine the congruency between the planned and actual activities. It includes three strategies: "The first is to detect or predict defects in the procedural design or its implementation stage, the second is to provide information for decisions, and the third is to maintain a record of procedures as they occur."[108] To deal with program defects, educators must identify and continually monitor potential sources of project failure. They must attend to the logistics of the entire operation and maintain communication channels among all affected parties. The second strategy involves decisions to be made by project managers during project implementation. For example, managers may decide that certain in-service activities are needed before program implementation. The third strategy addresses the main feature of the project design—for example, the particular content selected, new instructional strategies, or innovative student-teacher planning sessions. Process evaluation occurs during implementation. It is a piloting process conducted to debug the program before districtwide implementation. It enables evaluators to anticipate and overcome procedural difficulties.

Product evaluation has evaluators gathering data to determine whether the final curriculum product now in use is accomplishing what they had hoped. To what extent are the objectives being met? Product evaluation provides information that enables evaluators to decide whether to continue, terminate, or modify the new curriculum. For example, a product evaluation might furnish data showing that a science program planned for talented science students has allowed students to achieve the program's objectives. The program is then ready to be implemented in other schools within the system.

Humanistic Models, Postmodernist Models

Stake's and Stufflebeam's evaluation models draw heavily on the quantitative-technical approach to evaluation. Their models are most useful for addressing the standards and accountability demands of this century. They certainly find acceptance within the camps of cognitive science, educational psychology, computer science, and now neuroscience.[109] Also, their scientific models mesh with the thinking of those managers of the marketplace as well as of most politicians.

However, there seems to be a constant, but small, number of educators who believe that evaluators have bought excessively into the "education as a business within the marketplace" paradigm. Some educators have become mesmerized by observing or measuring the attainment of specific "learnings." They have spent excessive amounts of time generating elaborate evaluative schemes to measure program success.

Challenging this business posture, some educators are advocating more humanistic (naturalistic) or postmodernist methods of evaluative inquiry. These evaluators realize that actual learning is messy. Students and teachers are unpredictable actors in educational theater.[110] Individuals have different values, abilities, interests, dispositions, histories, cultures, and even different perceptions of reality. There are no standardized students. Thus, these evaluators argue for a more holistic approach to evaluation, one that provides detailed portraits of the situations being evaluated.

Evaluation reports are less lists of numbers than written descriptions of findings or occurrences. The approach focuses more on human interactions than on outcomes and more on the quality than the quantity of classroom or school life. Humanistic evaluators delve into the *why* behind the *what* of performance. The stress is on interpretative understanding rather than objective explanation.[111]

Whereas scientific evaluators might simply ask what students learned, humanistic, postmodernist evaluators query the value of the knowledge learned. These evaluators generate questions that cannot be answered with any finality.[112] Their questions produce responses enriched not with certainty, but with "difficulty, risk, and ambiguity."[113] The responses trigger in both the asker and the responders a myriad of moods and a universe of emotions.[114] Such questions are anathema to scientific, modernist, evaluators. Often, humanistic, postmodernist evaluators raise questions in their approaches that may not even relate to the aims of education. They realize in assessing the curriculum that it exists within political, social, and moral realms. Data must be processed as to its significance. Humanistic, postmodernist evaluators are cognizant that inquiry is not value-free. Even objective data exist within a sphere of subjectivity.[115] This acceptance of subjectivity allows focus on the true, the good, the beautiful, the just, the right, the spontaneous, the awesome, the amazed, the unexpected, the imaginative, the unique, and the emotional.[116]

EISNER'S CONNOISSEURSHIP AND CRITICISM MODELS. Elliot Eisner has recommended two humanistic evaluation models—connoisseurship and criticism—that draw heavily from the arts. Both models are designed to produce a rich description of educational life as a consequence of new programs.

Eisner describes *connoisseurship* as a private act engaged in to personally "appreciate the qualities that constitute some objects, situation, or event."[117] Connoisseurship has essentially five dimensions: (1) intentional, (2) structural, (3) curricular, (4) pedagogical, and (5) evaluative.[118] These dimensions reflect different aspects of curriculum and evaluation. *Intentional evaluation* refers to a personal assessment of a curriculum's value, merit, and worth. *Structural evaluation* assesses the curriculum's design and the school's organization. (According to Eisner, the spaces within which educators and students function influence the quality of the curricular experience.) *Curricular evaluation* assesses a curriculum's specific contents and how they are organized and sequenced. *Pedagogical evaluation* assesses instructional design and teaching strategies. (Does the instructional approach suit the curriculum's aims and content?) *Evaluative evaluation* assesses evaluation itself. How are evaluative data obtained? How is the curriculum assessed? Are tests and other evaluation methods giving a full and accurate picture of student progress?

The data sources for connoisseurship evaluation are many.[119] Evaluators observe teachers in the classroom and note how they interact with students. Evaluators might also interview students. Other data sources include the particular instructional materials used, student products, and teacher-made tests.[120]

Unlike connoisseurship evaluators, criticism evaluators share their critique of a new curriculum with the public. They interpret and explain the results of the new program. *Criticism evaluation* entails (1) description, (2) interpretation, (3) evaluation, and (4) thematics. Evaluators (1) write reports in which they describe the curriculum and educational environment; (2) interpret their findings for audiences—for example, by answering questions as to the reasons for the new curriculum; (3) attempt to determine and communicate the new program's educational value; and (4) ascertain from looking at the curriculum what theme or themes emerge. In considering

specific curricular situations, criticism evaluators seek to extrapolate general themes about learning and meaningful knowledge—themes that can guide curriculum development and execution.

By definition, connoisseurs possess expert knowledge. Educational connoisseurs must have knowledge of curriculum and instruction to determine what to observe, how to see, and how to value or appreciate. Good critics are aware of and appreciate a situation's subtleties; they can write about nuances in ways that help others become more aware of the phenomenon under consideration.

Eisner would have evaluators engage in qualitative activities—for example, participate in the classes they observe and ask many questions about the quality of the school and the curriculum. Evaluators following Eisner's model engage in detailed analyses of pupils' work. They use films, videotapes, photographs, and audiotapes of teachers and students in action. They note what is said and done, but also what is *not* said or done. They strive to describe the *tone* of the curriculum in action.

Eisner makes the point that evaluation should include reporting to the public (parents, school boards, local or state agencies, and so on). Evaluators must communicate the educational scene.

Slattery, in discussing the connoisseurship and criticism models, characterizes Eisner as a transitional figure moving away from modernism and toward postmodernism. Slattery purports that Eisner's models will be deconstructed by the postmodernists, revealing not a precise notion of expertise or masterpiece but templates echoing a multitude of voices and subcultures.[121] If we accept Slattery's judgment regarding Eisner, we might have to put all of the humanistic, postmodernist evaluation models in the transition realm. We further retort no one in the postmodern universe can say with any certainty that they are deep within the postmodern cosmos. For we do not know its dimensions, and if we did glance at them, we would realize that they are dynamic and ever changing; they are complex and chaotic.

ILLUMINATIVE EVALUATION MODEL. Another humanistic, postmodernist approach to evaluation is illuminative evaluation, sometimes called *explication.* Originally developed by Malcolm Parlett and David Hamilton, this approach illuminates an educational program's specific problems and unique features. To determine these problems and features, we must focus on the educational environment within which a curriculum is developed and delivered. Curricula rarely (if ever) are implemented and maintained as originally conceptualized and created.

Illuminative evaluation allows evaluators to discern the total program as it exists and functions and to gather data about its particular workings. The evaluator determines the results of the taught curriculum and identifies assumptions evident in its delivery; the attitudes and dispositions of teachers, students, and the public; and the personal and material factors that facilitate or impede the program.

Illuminative evaluation has three steps: observation, further inquiry, and explanation.[122]

1. *Observation.* Evaluators get an overview of the program and describe the context within which the curriculum is being delivered, considering all factors that might influence the program. They can gather data on the arrangement of school subjects, teaching and learning styles evident, the materials being used, and the evaluation methods employed by the teacher.
2. *Further inquiry.* Evaluators separate the significant from the trivial and seek to determine whether the program works and why it works or does not work. They gain a sharper focus from continually examining the program in action, spending extended time in the field. They also gather data by examining school documents and portfolios of students' work and by interviewing or giving questionnaires to staff and parents.
3. *Explanation.* Evaluators who use this model are not attempting to pass judgment on the program but to furnish data on what is happening with the program and why. Evaluators' explanations are presented to the people affected by the program, who then make decisions.

The illuminative approach is holistic and subjective. Observed interactions are not broken down into discrete categories for measurement, but considered within the context of their environment.

Action-Research Model

Action research is an evaluative approach that blends the scientific, modernist and humanistic, postmodernist. It is concerned with continual modification of the educational experience so that every educational event is fresh.[123]

Action-research evaluation is distinguished by direct participation in the curriculum. Parker Palmer states that the only way to evaluate teaching and learning is to be present within the learning environment.[124] Teachers are the key players in action-research evaluation. They evaluate both the curriculum and the teaching of the curriculum. They are willing to take chances and learn partly by trial and error.

When the action-research approach is weighted toward research, evaluators investigate quantifiable results of particular classroom actions—results that they hope will allow them to generalize to similar groups of students in similar classrooms. The data suggest general approaches to creating and delivering curricula. They also encourage self-evaluation by teachers and provide insights into the effects on teachers of conducting research within their classrooms and schools. Such data illuminate how teachers' attitudes and prejudices affect student learning.

When action-research evaluation is weighted toward assessment, it is not concerned with education in general but with the unique classrooms of individual teachers. It does not focus on gathering data from which to generalize to other teachers, students, and classrooms. It is concerned with engaging a specific teacher in problem solving to optimize the learning of specific students at a particular time. Gathered data are used to determine whether to continue or modify a particular curriculum or particular instructional approach. The teacher continuously adjusts content, teaching, and educational experiences.

The first step in this fine-tuning is for the teacher to identify what he or she wants to accomplish with a particular aspect of the curriculum or a particular pedagogy and what students wish to accomplish from their engagement with the curriculum. The next step is to determine how to monitor the implemented curriculum. The third step is to interpret the data gathered during monitoring. The fourth step is to continue the process of action research. This step can be accomplished only by teachers who gather data during the actual teaching of the curriculum. Teachers may videotape their teaching, have colleagues observe their teaching, take time from their teaching to record actions and their results in journals, interview students after a particular educational activity, and of course, administer tests.

Figure 9.3 depicts the general sequence and feedback of action research. Table 9.1 provides an overview of evaluation models.

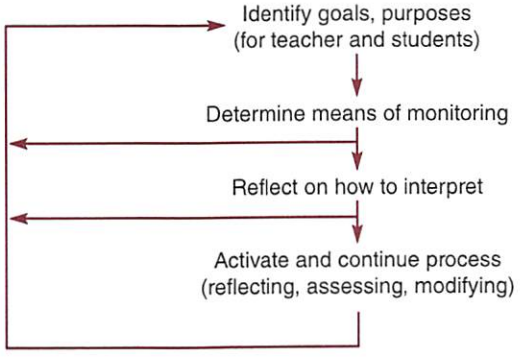

FIGURE 9.3 General Sequence/Feedback: Action Research

Source: Based on the comments by Collin J. Marsh and George Willis, *Curriculum: Alternative Approaches, Ongoing Issues*, 4th ed. (Upper Saddle River, NJ: Pearson, 2007).

Table 9.1 | Overview of Evaluation Models

Model	Author	Approach	View of Reality	Possibility of Generalization	Role of Values
Congruence-contingency	Stake	Scientific, modernist	Reality is tangible, single.	Yes	Value free
Context, input, process, product	Stufflebeam	Scientific, modernist	Reality is tangible, single.	Yes	Value free
Connoisseurship/ criticism	Eisner	Humanistic, postmodernist	Realities are multiple, holistic, ever changing.	No	Value bound
Illuminative	Parlett and Hamilton	Humanistic, postmodernist	Realities are multiple, holistic, ever changing.	No	Value bound
Action research	Wolf	Humanistic, postmodernist, Scientific, modernist	Realities are multiple, holistic, ever changing.	No Yes	Value bound Value free

TESTING

> This is an age of examinations. . . . Is it not a wonder that so many of our American boys and girls survive the almost continual examinations to which they are subjected? There are oral examinations, written examinations, daily examinations, weekly examinations, monthly examinations, quarterly examinations, yearly examinations, examinations for admission, examinations for promotion, examinations for graduation, competitive examinations . . . in short, there is no end of examinations in this life.[125]
>
> —Charles I. Parker (high school principal, Hyde Park, Illinois), 1878

> However much we may deprecate the evil of cramming and other mis-directions of energy, and deplore its waste, . . . it must be admitted [that mandatory and demanding examinations] mean the thorough awakening of the schools.[126]
>
> —L. E. Rector (educator from Jersey City, New Jersey), 1895

The above quotes hopefully provide you, the reader, with a sense that testing in the United States has an extensive history. Throughout this history, we have had advocates of increased testing and critics such as Parker who indicate that we are caught in a testing tempest. Today, we are still in a maelstrom of debate about testing and holding schools accountable for highly educating their students.

As William J. Reese asserts, written examinations have become well established within our educational system. The expanding complexities of our world insist we respond to these dictates and furnish evidence that our educational actions are effective. We must assess whether we are providing relevant curricula and effective pedagogies to meet not only economic demands, but also social demands. In fact, we frequently feel that the public is making ultimatums that schools produce renaissance individuals. This utopian aim is not possible, even if we started formal education of children from birth. If attempted, we would have only 18 years at a minimum and 22 years if including college. And, as Reese notes, even if such a person could be nurtured and developed with advanced degrees and postdoctoral study, there can be in the 21st century no guaranteed economic opportunities.[127] Additionally, psychometricians have yet to develop a test that measures one's understanding of knowledge yet discovered or formulated. Tests cannot accurately measure students' aptitudes for occupations not yet envisioned.

Testing, while in constant debate, is well situated in this country and its schools. Testing is big business. Reese notes that the Educational Testing Service in Princeton, New Jersey, is the largest "nonprofit" business there is. The company develops in excess of 50 million tests annually for over 180 countries. It further manages the taking of the tests and scores such tests.[128] Even educational publishing firms are entering test development and administration. Also, there are many businesses engaged in educational tutoring to prepare students for these exams.[129]

The test has been pervasive through much of our history. It is even more pervasive in this new century attempting to define our relationship to questions of trust, knowledge, and even reality.[130] We seem continuously poised as a society to blame the schools when society in general has problems. Politicians often ignite a distrust of schools and a dissatisfaction with the quality of curricula and teaching. No Child Left Behind was created by politicians, not educators. A Nation at Risk was a political critique of the American educational system. Race to the Top is a political animal based largely on myth regarding American schools.

David C. Berliner and Gene V. Glass have written a book titled *50 Myths and Lies that Threaten America's Public Schools*. Myth 1 is that international tests reveal that U.S. schools produce a second-rate education.[131] Such comparison reads too much into a score. Berliner and Glass also point out that the United States is much more heterogeneous than other industrialized nations. You need a lot more information in order to determine who has the better school system. And, as they posit, "better for whom? better on what criteria?"[132]

High-Stakes Tests

> [H]igh-stakes testing is taken as an a priori assumption in educational policy. An educational system without high-stakes testing is nearly unthinkable, unimaginable, . . . the tests are "here to stay."[133]
>
> Teaching and assessment play critical roles in helping students develop an understanding of why they study different subjects in school. . . . The purpose of learning" is "to do well on tests."[134]

Wayne Au denotes that a test is high-stakes when the information it provides is employed in making important decisions that impact all the educational players directly engaged in teaching and administering a school. Also, the data can influence the entire school district and the community itself.[135] High-stakes tests determine whether a student graduates from high school. Such tests can be used to decide teachers' and administrators' salaries.[136]

Education is expensive. The public is increasingly concerned with getting the most for their money. The public demands that schools maintain high academic standards. Certainly, every person wants the best that can be obtained. Parents realize what their students learn in school will contribute in important ways to their future successes. Schooling and education are integral parts of a high-stakes life game.

National associations of content and discipline specialists have created standards that give consideration to student knowledge of specific content, skills, and procedures. The standards of the National Council of Teachers of Mathematics, National Research Council (which sets science standards), National Council for the Social Studies, and National Council of Teachers of English have significantly influenced assessment. State departments of education, as well as most U.S. school districts, have taken note of these standards and the public's demand that they be met. These standards are provided to guide teachers' curricular and instructional actions and influence the performance levels that students must demonstrate.

However, are standards to be used as guides? Increasing numbers of educators are perceiving standards not as guides to teachers' and students' actions, but as controls and regulations of what occurs regarding curricula and instructional strategies. High-stakes standardized tests are being used as instruments to determine how close educators and students adhere to the standards most frequently set from afar. If students and teachers miss their marks, they are penalized. Students may not be advanced or get diplomas, or teachers may not have contracts renewed. Schools can even be shut down.

Au notes that with the emphasis on high-stakes testing, there is a narrowing of curriculum content. Content is selected to match what is on the test. Essential subjects are only those that are tested. Subjects considered nonessential receive less emphasis or are eliminated. Many schools have reduced or eliminated subject areas such as art and music. Some schools have even eliminated recess—it's not on the test. Physical education usually is not part of the high-stakes testing picture.

Au suggests that high-stakes testing controls not only the content, but also the manner in which content is experienced. Teaching to the test shapes curriculum form—"the organization

of meaning and action, including the order in which [students] are introduced to content and the very form that knowledge itself takes, in the curriculum."[137] The flow of knowledge organization suffers as the content knowledge is dissected so that it meshes with how the high-stakes test will measure students' mastery of it.

Not only is the content being molded and organized to mirror what is contained in the high-stakes test, but teachers are having to relinquish their instructional strategies and accept those pedagogies that correlate "to the forms of knowledge and content contained on the high-stakes tests."[138] Some assert that teachers are abandoning what they consider best practice in order to be in compliance with standards-based education and to be judged accountable.[139]

Standards from professional and other organizations, both public and private, have certainly increased testing in public schools. Currently, there is considerable controversy regarding the soundness and consequences of testing to address particular standards. Do we want to narrow the curriculum? Do we desire to shape how the curricular content is organized? Do we wish to limit the creativity of teachers in the way that they orchestrate their instruction? Finally, do we want various outside sources at local, state, and even federal levels essentially to determine school policies with regard to curricula, instruction, and approaches to evaluation?

It appears that we do, or at least, that educators are not able to counter the demand for being accountable and efficient. Of course, educators do wish to be accountable; they wish to be effective in educating their students. However, are the key criteria for evaluating education efficiency to be the maximum amount of content knowledge learned in the least amount of time and the speed in which skills are demonstrated at high rates of accomplishment? As Taubman asserts, it does seem that testing, especially high-stakes testing, is now defining not only our approach to education, but just what we mean by student and teacher "knowing" and competencies.[140]

Today, all states have statewide testing programs. Vast numbers of school districts have their own districtwide testing programs. Testing, it seems, is almost the school's major educational activity. Often, as suggested before, whether students advance to the next grade or graduate depends on whether they pass or fail a particular test.[141] Teachers whose students pass such exams tend to be evaluated more favorably than teachers whose students fail. Some people, as indicated earlier, recommend that teacher pay should be determined by the performance of their students on these high-stakes tests. Pay for performance has been in the news for more than a decade. Matthew Springer and Catherine Gardner note that Google News reported in 2010 that an average of 4,558 news stories per year dealt with teacher pay being determined by student performance on tests.[142] States such as Texas, Florida, and Minnesota have allocated more than half a billion dollars to incentive pay programs that aim at rewarding teachers for "effective" teaching. The funding of the federal Teacher Incentive Fund was quadrupled in 2010. The Race to the Top federal program emphasizes performance pay. This program has allocated more than $4 billion to this effort.

It does appear, as Springer and Gardner assert, that pay for performance is poised to become a reality factor when evaluating educational effectiveness.[143] This being the case, educators and those advocating for increased effectiveness of education must query themselves about how we are to define teacher and student performance. Certainly, one score on a high-stakes test cannot be the sole deciding indictor. As Taubman cautions, "in reducing everyone and everything to quantifiable data from test scores and attendance records to performance on behavioral check sheets, all historical, personal, idiosyncratic, and context-specific details about the person or event are erased, creating, as the anthropologist Geoffrey C. Bowker states, 'the least possible information that can be shared about events, objects, and people while still maintaining a viable discourse around them.'"[144]

We mentioned in a previous chapter that when standards are emphasized at the time of creating aims, goals, and objectives, there is a tendency to engage in activities that standardize the educational experience of both teachers and students. We cited some cautions. Taubman states that in enforcing standards and a standardization of curricula and instruction, we endanger individuals' idiosyncrasy. In using the same metric to measure "attainment of the standard," we break down human spirits and behaviors into a sameness that crosses boundaries, both geographic and intellectual.[145] Employing the same metric ignores that students are diverse, unique, and differing in abilities, interests, values, beliefs, anxieties, dispositions, and often language.[146]

9.2 Narrowing the Curriculum in School
Critics of high-stakes testing complain that schools now focus almost exclusively on tested subjects like language arts and math. Watch this news segment on the narrowing of the curriculum. In this age of high-stakes testing, how would you evaluate the current curriculum in your school district?

https://www.youtube.com/watch?v=VxOVvxogpt0

High-stakes testing has caused many teachers to game the system—not only teaching to the test, but coaching students with sample test questions or even excusing those students who might not do well on the test to have a "day off." Although gaming the system might boost test scores, are such scores evidence of high-quality learning? Indeed, that is the key question with regard to all tests, either teacher-made or standardized. What do the resulting scores actually tell you? According to Alfie Kohn, tests, especially standardized tests, provide scant information about what students actually know and can do. Tests can indicate that some students are more proficient than others, but we still do not know how proficient each student is regarding specific subject matter.[147] Likewise, tests can indicate that one teacher's students attained higher scores than another teacher's, but the scores do not note with any precision whether one teacher was more effective than the other.

It appears that most tests administered by U.S. schools measure knowledge in an unsophisticated way. Various studies have indicated these tests require of students only relatively shallow thinking.[148] Essentially, they test superficial knowledge, not understanding.

Norm-Referenced Tests

Norm-referenced tests (NRTs) are the most commonly used. A student's performance on a particular test is compared with that of other students who are his or her peers. The items in an NRT usually address a wide area of content. The students, as a group, establish a norm. Students can be grouped by age, grade level, ethnicity, sex, geographic location, or any other easy-to-categorize factor. In order to make comparisons among the students, these tests must be administered to the students in similar fashion and formats and at basically the same time. The manner of scoring the tests must also be the same to furnish meaningful comparison data.[149]

Standardized achievement tests are probably the best-known NRTs. They provide information useful in ranking individual students or groups of students. Specifically, these tests identify which students are successful in their learning and which students might require remediation. Are the students who took this test progressing at a rate comparable to their peers? If groups of students are tested only once, the test results have questionable value for measuring the quality of a curriculum or instruction. However, when such tests are administered each year at the same time, then the test data can furnish information that depicts patterns revealing both the quality and shortcomings of the curriculum and instructional strategies.[150] However, teachers must realize that NRTs do not specifically relate to a particular curriculum, nor do they effectively measure what has been taught. They do not indicate what a student can or cannot do, nor do they provide evidence that a student knows or does not know specific content.[151] In addition, many educators fail to realize that different standardized achievement tests are not interchangeable.[152] When educators use a particular test to rank their students with regard to other students who have taken a different standardized achievement test, the rankings cannot be accepted with any confidence. When states employ such tests to compare their students with students in other states, they cannot reach meaningful conclusions regarding the relative worth of their curricula.

W. James Popham faults the educational community and the general public for ignoring the nature of standardized tests used in curriculum comparisons or various other educational research attempts. He states that "inadequate scrutiny of the tests used in key investigations is particularly galling whenever a study's results indicate that there is 'no significant difference' between the achievement of students from one group to that of another group."[153] He indicates that reporting no significant difference deprives us of any useful conclusion. Standardized achievement tests cannot detect the "differences between students taught effectively and students taught ineffectively."[154]

Research indicates that standardized achievement tests highly correlate with students' socioeconomic status. This high positive correlation obscures the impact of educational efforts such as new curriculum. Despite these limitations, educators continue to employ standardized tests to determine the curriculum's success and evaluate teachers' effectiveness. Educators

continue to use such tests to rank students in various schools and to determine which students should advance or graduate.

Criterion-Referenced Tests

The most common alternative to the NRT is the criterion-referenced test (CRT). The CRT is designed to indicate how a student performs a skill or task, or understands a concept, with respect to a fixed criterion or standard. The performing of a skill or task is measured against what are defined as proficiency or achievement standards. The depth of understanding of a concept or certain content is measured by a content standard.[155]

Currently, many of these standards are created by groups outside of school districts (state education agencies or state legislatures). Often, the standards are broken down into specific objectives, frequently stated in behavioral terms. For example, a CRT might require a learner to identify longitude and latitude lines on a map or to multiply two-digit numbers. Well-delineated descriptions of the learning are the key features of such tests. This specificity enables educators to determine precisely what a student does or does not know—or can or cannot do—in relation to a particular curriculum. The score on each item interests the evaluator. The teacher wants the student to master the content, skills, or attitudes addressed in each item. Teacher and student will persevere until the student gets the test item right.[156]

CRTs indicate changes in learning over time (in contrast, NRTs measure learning at a specific time). As Taylor and Nolen indicate, teacher-made tests most often are CRTs administered to determine the proficiency of a student's learning in relation to a standard or goal.[157] For CRTs to indicate student mastery, the criterion must be appropriate. Most educators consider 80 percent correct as indicative of mastery. Why? We don't know exactly, but 80 percent does seem to indicate a high level of performance. However, we must consider a test item's age-appropriateness. Otherwise, a test item might be so easy that everyone scores 80 percent or higher, or so difficult that no one does.[158] We must also ask ourselves if a standard of 80 percent is appropriate for all learners in all realms of the curriculum. A level of 80 percent mastery might suffice with regard to understanding a book but not suffice with regard to conducting a science experiment. Likewise, 80 percent is inadequate with regard to accounting exercises (which require 100 percent accuracy).

W. James Popham notes that when educators employ criterion-referenced tests, they need to consider what is an optimal grain size. He defines grain size as "the breadth of a criterion domain."[159] We would add, must all students attain similar or identical grains in all subject areas where criteria have been identified? In raising this query, Popham is viewing a criterion not as a level-of-performance, but criterion-as-domain. He notes that while assessing student performance is important, the purpose of criterion-referenced measurement criteria is to specifically delineate the skills or knowledge being assessed.[160] We assert these tests do both indicate a level-of-performance of specific skills and curricular content. Popham cautions that if the grain size of contents and skills is too narrow or too vast, it will not be of value in assessing the effectiveness of pedagogies or curricula.[161]

The grain size essentially deals with the level of specificity. If the specificity of CRTs is intense, as noted previously, it can be a disadvantage. Because such tests address specific objectives, as many as 10 to 15 tests may be necessary to get a thorough picture of the curriculum.

The primary value of CRTs is that they are curriculum specific. They enable curriculum evaluators to assess a new curriculum in their school districts. Evaluators also can determine the instructional realm's effectiveness and whether certain content and skills have been taught. The tests are good tools for assessing student learning and teachers' pedagogical approaches.

It is not easy to determine the standards for acceptable performance. Just what is the cutoff score for mastery of an objective? Educators usually set the passing score somewhat arbitrarily. Perhaps the most serious criticism of CRTs is that most lack information regarding their reliability. In fact, most are constructed without any attention to reliability. However, CRTs have curricular validity: The items usually coincide with the curriculum's objectives.[162] Table 9.2 presents a comparison of NRTs and CRTs.

Table 9.2 | Comparison of Norm-Referenced Tests (NRTs) and Criterion-Referenced Tests (CRTs)

Characteristic	NRT	CRT
1. Comparisons made	Score to group average	Score to minimum standard
2. Purpose	Survey or achievement test	Mastery or performance test
3. Validity	Content, criterion, or construct	Content *and* curricular validity
4. Degree of validity	Dependent on instruction	Usually high
5. Reliability	Usually high	Usually unknown
6. Importance of reliability to test model	Important	Unimportant
7. Traits measured	Exist in varying degrees	Present or not present
8. Usability		
Diagnoses	Low general ability	Specific problems
Estimation of performance	Broad area	Specific area
Basis for decision making	How much was learned	What has been learned
9. Item difficulty	Medium	Easy items
10. Administration	Standardized	Variable
11. Size of group tested	Large	Small
12. Content covered	Broad	Narrow
13. Skills tested	Integrated	Isolated
14. Control of content	Publisher	Instructor or school
15. Limitations	Inability of school personnel to interpret tests on local level	Difficulty of constructing quality tests
16. Versatility	Extensive	Limited
17. Comparison of results between schools	Readily available	Not yet developed
18. Distribution of scores	Normal (one)	Rectangular (two)
19. Range of scores	High	Low
20. Repetition of test if test is failed	No, one test	Until mastery occurs
21. Basis for content	Expert opinion	Local curriculum
22. Quality of items	High	Varies, depending on ability of test constructor
23. Pilot testing	Yes	No
24. Basis of item quality	High discrimination	Content of items
25. Student preparation	Studying for test does not help much	Studying for test should help
26. Teaching to test	Difficult to do	Encouraged
27. Standards	Averages	Performance levels
28. Scores	Ranking, standard score, or number correct	Pass or fail
29. Type of measure	Relative	Absolute
30. Purpose	Ranking students	Improving instruction
31. Revision of test	Not possible	Often necessary
32. Student Information about test content	Little available	Known in advance
33. Motivation of students	Avoidance of failure	Likelihood of success
34. Competition	Student to student	Student to criterion
35. Domain of instruction	Cognitive	Cognitive or psychomotor

Source: Based on Allan C. Ornstein and David A. Gilman, "The Striking Contrasts between Norm-Referenced and Criterion-Referenced Tests," *Contemporary Education* (Summer 1991), p. 293.

Subjective Tests

NRTs and CRTs are both categorized as objective tests. This essentially means that the test questions have one correct answer. However, curriculum evaluators also have access to subjective (constructed-response) tests. These tests have many correct responses to each question. For this reason, they are much more challenging to score than objective tests. Often, it is the depth or creativity of the response that determines the evaluative ranking. Essay tests are subjective. Style, insight, originality, use of accurate information, strength of argument, and knowledge of the topic are criteria by which an essay is judged. If educators wish to use essay questions to compare students or programs, the essay questions presented must be the same for all students.[163]

ALTERNATIVE ASSESSMENT

Since the early 1900s, student data have been gathered by means of teacher-made or standardized tests. Today there is an increasing call for alternative forms of assessment.

States and school districts are engaged in efforts to better align tests and other evaluation efforts with state and district standards and to create means of assessment that truly capture students' knowledge and skills.[164] Many new forms of assessment involve open-ended tasks; students are required to use their knowledge and skills to create a product or solve a problem. Such evaluation events are called *performance assessments.*

Many educators consider performance assessment to be synonymous with authentic assessment. Certainly, they both are examples of alternative assessment, because they employ methods other than multiple-choice or like-developed objective tests. However, in 1992, Carol Meyer argued that performance assessment and authentic assessment are not the same. For an alternative (performance) assessment to be authentic, it must engage students in tasks and activities that resemble actions in the real world. The tests cannot be contrived by the teacher.[165]

A writing exercise is an example of a performance assessment, but it may not be authentic. For instance, here is an example of an inauthentic assessment of students' writing skills. The teacher presents the students with a precise formula for preparing to write and actually writing a short story. On the first day, the students have 50 minutes to generate the story's topic; on the second day, they have 50 minutes in which to create a rough draft; and on the third day, they have 50 minutes in which to revise and prepare the final draft.[166] Certainly, the students have been engaged in the writing process. However, actual writers do not follow such a restricted process in their writing of short stories. Thus, the contrived activity is not authentic. To make this writing of a short story more authentic, teachers might indicate that students should engage in creative writing throughout the year when the spirit moves them and then file such writing in portfolios. Students select the time for their writing and decide when to share their drafts with the teacher and other students. They revise their drafts according to their own schedules. In this case, students are engaged in an authentic writing assessment, writing in a way that resembles the way that professional writers actually work.

Authentic assessment includes real problem solving, designing and conducting experiments on real problems, engaging in debates, constructing models, creating videotapes of performances, doing fieldwork, creating exhibits, developing demonstrations, writing in journals, creating new products, formulating computer simulations, and creating portfolios. Authentic assessment employs strategies and approaches that present students with real-life situations and conditions.[167] Authentic assessment is more than the gathering of students' products. It involves teachers' observations and inventories of students' work with accompanying commentary regarding the judgments made. Authentic assessment reports on individuals and groups within the classroom.

Table 9.3 | Alternative Assessment versus Traditional Assessment

Alternative Assessment	Traditional Assessment
Samples: student experiments, debates, portfolios, student products	*Samples:* multiple-choice tests, matching tests, true-false tests, completion tests
Evaluation judgment based on observation and subjective, yet professional, judgment	Evaluation judgment based on objective recording and interpretation of scores
Focus on individual students in light of their learning	Focus more on score of student as it compares with scores of other students
Evaluator able to create an evaluation story regarding an individual or group	Evaluator able to present student knowledge as a score only
Evaluation that tends to be idiosyncratic	Evaluation that tends to be generalizable
Furnishes data in ways that allow curricular action	Furnishes data in ways that inhibit curricular or instructional action
Allows students to participate in their assessment	Tends to place evaluation under the aegis of the teacher or external force

Source: Based on Dennie Palmer Wolf and Sean F. Reardon, "Access to Excellence through New Forms of Student Assessment," in Joan Boykoff Baron and Dennie Palmer Wolf, eds., *Performance-Based Student Assessment: Challenges and Possibilities*, Ninety-fifth Yearbook of the National Society for the Study of Education (Chicago: University of Chicago Press, 1966), pp. 52–83.

Table 9.3 presents some comparisons between alternative, authentic assessment and traditional paper-and-pencil test assessment.

We believe that both alternative and traditional assessment should be used. Educators sometimes accept new practice too readily. Dennie Wolf and Sean Reardon caution, "If new forms of assessment are to work, they require serious gestation."[168] Educators must reconceptualize intelligence, rethink what it means to know something, redefine excellence, and rethink their measurement habits. At the same time, educators must be careful not to interpret the new means of evaluation with traditional mindsets.

William Glasser has proposed seven features of optimal assessment. First, assessment itself should foster student growth. Second, it should allow us to see the consequences of instructional effects. Third, assessment should illuminate the processes and products of learning. Fourth, it should involve student self-assessment; that is, students should be active participants in judging their achievement. Fifth, assessment should be an integral part of group activity. Assessment data should inform the educator not only about what a student knows, but also about how well the student works with others and adapts to group dynamics. Sixth, assessment should entail meaningful tasks that tie in to overall learning and the curriculum's knowledge goals. Seventh, assessment should be comprehensive, addressing a broad range of information and skills rather than centering on narrow understanding of a particular content.[169]

Alternative assessment should be an ongoing activity integral to curriculum enactment, not an activity engaged in only at particular times of the year to obtain information on student progress. Teachers and students should continually question how well things are being taught and learned. A paper trail should elucidate the quality of student learning.

New assessment methods require new assessment criteria. George Hein would include a moral standard among indicators of effective schooling. A school curriculum that meets a moral standard provides students with skills and knowledge requisite for contributing to the general social good. As Hein indicates, moral purpose was central to the progressive education philosophy.[170]

The portfolio is perhaps the most popular method of alternative assessment. Because it is a sampling of student work over time, a portfolio provides evidence of a student's understandings, skills, and behavioral dispositions. It often records a student's degree of effort and participation in learning. Taylor and Nolen identify several different kinds of portfolios, each one having a different purpose: showcase portfolios, growth portfolios, process portfolios, and cumulative portfolios.[171]

Showcase portfolios, truthful to their name, use concrete examples to emphasize what students have attained in a particular time capsule and at a particular level of accomplishment. Such a portfolio might show a student's art for a given year or samples of a student's essays. With regard to a science showcase, the portfolio might present write-ups of experiments done or notes on field studies.

Growth portfolios provide a visual mapping of a student's increased skills or competencies or understandings over time. A student, often assisted by a teacher, plots out way points denoting progress in both declarative and procedural knowledge. Such a portfolio serves to both guide and inspire students in their learning journeys. For example, a portfolio can include a composition that a student wrote at the beginning of the school year and another composition written at year's end. The student and teacher can critique the two papers to determine writing progress. As Taylor and Nolen denote, growth portfolios enable students to assess their increased competencies regarding learning a completely new subject or skill. Such a portfolio is most informative to students and teachers in activities such as learning a new language.[172]

A useful device for documenting students' process or enactment of procedural knowledge is the process portfolio. Here materials included denote how successful students have been in accomplishing authentic performance. These authors define *authenticity in work* as that which has relevance and authenticity in the out-of-school world.[173]

The fourth type of portfolio, the cumulative portfolio, is part of the summative evaluative data story. This portfolio contains a student's entries of all his or her work for a year or even longer. The works presented are considered by both student and teacher to be the best examples of work done and tasks attained. Taylor and Nolen posit that these cumulative portfolios become part of students' cumulative records, denoting their progress during their total school experience. Teachers at the start of each year can use cumulative portfolios from the previous years to personalize the curriculum for the incoming students.[174]

For students to create each type of portfolio at a quality level, they must determine, with the teacher's assistance, what criteria are to be employed in judging what should be included. Of course, as students progress through the year or years, they can delete material that, upon later reflection, does not exemplify quality work. Specifically, students working with teachers must engage in critical analysis of their work and their learning strategies.

One of the greatest benefits of portfolios is that students are major players in their own evaluation. Students must reflect on their work, critique their level of understanding, and judge their study and analytic skills. Portfolios enable, even demand, that students continually self-evaluate, not for a grade, but to increase the quality of their learning procedures, the depth of their understandings, and the significance of their resultant works. Additionally, students can utilize this alternative evaluative instrument to personalize their curricular experiences.

Portfolios essentially allow students to present themselves as whole individuals. Portfolios enable students to become their own scholars and to define their works in regard to value and significance. In using portfolios, students use their voice to add evidence regarding their progress. Students have more than a list of scores or letter grades. Portfolios furnish students, teachers, and parents with material for conversation.

HUMAN ISSUES OF EVALUATION

We are not widgets. Wayne Au counters that the way we engage in evaluation assumes we are.[175] Yet despite our evaluative actions, deep down we realize that we are individuals with diverse personalities, talents, dispositions, interests, values, emotional stabilities, and intellectual

capacities. This section deals with human issues of evaluation, yet the human dimension seems absent in our evaluative deliberations and actions.

Students have been quantified, objectified, made into commodities to be molded, assembled, inspected, and then compared in the world marketplace.[176] We have standardized them and embraced the assumption that all students are essentially alike. We have touted that our tests are indeed objective and that factors such as local cultures, ethnicity, languages spoken, racial group, and socioeconomic status are essentially meaningless.[177] We need not really consider the environments of the "factories" in which the widgets are manufactured. All we must do is measure and judge the quality and quantity of the widgets.

However, ignoring the environments within which evaluation occurs means that it often is destined to fail, despite being valid in all technical details. Evaluation must be sensitive to ethnic or racial bias. Evaluation must be enacted with sophisticated consideration of the evaluative process and the social milieu. Evaluation is shaped by the stakeholders to whom it is reported. Evaluation neglecting the manner of presentation risks having evaluation results being misused, misinterpreted, or simply ignored.

Today, there exists a hidden dimension in evaluation activities: control. This control is over the teachers, the students, the curriculum. A central question is, Who is behind this control? We know that evaluation entails value judgments. The key question is, whose value judgments are they, and are they worthwhile? Who is deciding the purposes of education and the standards by which education is judged? There is no definitive answer. It depends on the sociological nature of various communities. However, it is apparent that evaluation is part of the political process. Often, schools release test results not to improve programs, but to please various power groups within the community or demonstrate to legislators that an educational program is effective. Sometimes test results are broadcast to convince various minority groups that their children are experiencing equity within the school system.

Not only that, students are being tested, via standardized testing, with fairness and equality. All students are being measured with the same metric. The tests put everyone on a level playing field. However, we assert that although the standardized tests may put students on a level evaluative playing field, they certainly ignore a level playing field when it comes to instruction and the curriculum. When evaluating students, we must consider their social, economic, ethnic, and, certainly, educational backgrounds. Not all students come to school with equal backgrounds for assuring school success.

We agree that our students function in a society that celebrates meritocracy. Accepting this, most citizens believe that "regardless of social position, economic class, gender, or culture (or any other form of difference), based on merit and hard work, any individual competes freely and equally with other individuals in order to become 'successful.'"[178] Yet such a belief is challenged by reality. Clearly, some students come to schools more likely to succeed in our schools and to master the curriculum. We do have inequalities in our society that place many students at a disadvantage regarding school success. As previously mentioned, even test designers bring various assumptions about social, cultural, and ethnic groups to their test making. If certain minority students do *too* well compared with the dominant white group, test items are not used in the next designed test. Experimental test items have to reflect what psychometricians assume about our diverse populations. And there is considerable debate if we should use our schools to reproduce our current society. Certainly, we must consider multiple interests when designing curricula and creating evaluative measures to determine student success. However, many educators and evaluators are reluctant to confront issues such as social justice within the educational system.[179] Educators are eager not to polarize communities and stir up controversy. However, fairness is crucial to consider when dealing with evaluation.

According to James Pellegrino, Naomi Chudowsky, and Robert Glaser, the idea of comparable validity is at the core of fairness. A fair test furnishes data from which we can draw valid inferences across individuals and groups.[180] Many people believe that tests tend to be

biased in favor of students who belong to the dominant culture. Tests use language and terms more familiar to the mainstream majority than to minority cultures. Students bring their cultural backgrounds and world knowledge to test situations. Deborah Meier states that "any choice of subject matter, vocabulary, syntax, metaphors, word associations, and values presupposes a certain social and personal history. We may have equally big vocabularies, but different ones. We may be speaking a grammar that is consistent and accepted, but not the standardized one used in academia."[181]

Evaluators and test designers realize that certain test items produce different results among students from different groups, even when all students have been matched in ability regarding the attribute or knowledge being assessed. For example, students responding to a test question about the discovery of the Americas might well respond differently depending on whether their cultural group sees the actions of Europeans as discovery or conquest. Students raised on farms are more likely than inner-city students to answer a question about agriculture correctly.

Also, is it fair to hold students with disabilities to the same standards as other students? Obviously, wheelchair-bound students cannot be held to physical education standards. Should students with reading and writing disabilities have to meet school standards in order to advance to the next grade or graduate?[182] Should we furnish computer systems to students with reading disabilities to help them with their reading?

The issue of fairness also affects evaluation of students classified as gifted. How do we judge the performance of such students? Many secondary students in advanced-placement and college-level classes complain that their A's look no different on their transcripts than the A's of students in the regular curriculum. Is this fair?

Evaluators are attempting to address the issue of fairness in evaluation by looking at a variety of means of evaluation. Certainly, the alternative methods of evaluation are useful here. Also, we can have grading that is based on multiple criteria. Several evaluators and assessment experts suggest that to really address the issue of fairness, we must consider students' backgrounds when we engage in evaluation. If we do this, we will be able to make conditional inferences from the data analyzed.

Students experienced with particular tasks find them easier than inexperienced students of similar innate ability. Confronted with a new aspect of the curriculum or a new problem, students first determine whether they have background information on which to draw. Those who do are likely to deal successfully with the content or problem. Those who don't may find the content or problem beyond them. We cannot simply say that some students succeeded and others failed. We must consider students' background when we make an evaluative judgment.[183]

Evaluation should encourage, not intimidate, students. It should foster cooperation and a sense of community among students rather than feelings of tense or aggressive competition. Teachers should present tests as learning experiences, not as means of reward and punishment. Much evaluation, especially standardized testing, produces fear among both students and teachers. Deborah Landry investigated the behavior of 1,058 K–5 students during a standardized reading test by asking teachers to report on their observations of these students. Landry conducted an online survey of 63 teachers and interviewed four others. The teachers reported that the standardized testing produced anxiety in the students, who commonly sighed, moaned, and even cried. Teachers reported that 49 percent of the students fidgeted during the testing; 33 percent were worried about how hard the test was; and 21 percent said they were nervous. Landry concluded that the students' behavior indicated strong feelings of helplessness, fear, abandonment, and self-doubt.[184] Other studies of standardized testing have yielded similar results.

Must we test all things? It seems so. As Landry reported, we seem to be not only assessing our students, but also creating psychological problems in students that we are not

assessing. In some cities, infants are being assessed as to whether they will fit in to particular preschools or kindergartens. In 2006, Peg Tyre wrote an article querying whether, in the first grade, we are doing too much too soon.[185] Must first graders be tested for everything? Must we score their play? Must students measure up right from the beginning of school? Where is the emphasis on the uniqueness of individuals?

It does appear that we are evaluating with such frequency and intensity that we are smothering students' joy of learning. With our push for standardizing, students are becoming widgets to be shaped and polished. Even students who are precocious are not always ready to be evaluated and sorted by psychometric devices. Tyre noted in 2006 that it appears that early schooling has become "less like a trip to Mr. Rogers' Neighborhood and more like SAT prep."[186]

Evaluation of students during their schooling has become excessive. We probe, poke, measure, assess, judge, sort, encourage, and discourage students so we can inform them as to how they measure up with regard to others. Educators should not make evaluation like a gauntlet that students must somehow survive. The educational experience should not be a series of pressured encounters for grasping a brass ring.

CHALLENGES IN THE 21ST CENTURY

In this chapter, we have focused on the approaches to evaluation, the mechanics of engaging in curriculum evaluation, and the types of tests and various assessment procedures. We have noted that we utilize evaluation to judge our curricula, our pedagogies, and students' learning. We have essentially followed a safe route, not being specific about what particular contents and what pedagogical strategies should receive our evaluative efforts.

But in the 21st century, this chaotic, complex, and morphing time, we need to evaluate what we are stressing in our educational programs and certainly shaping by the tests we construct. We certainly are stressing science, technology, engineering, and mathematics (STEM). And some have added the arts. Howard Gardner in 2011 published a book titled *Truth, Beauty, and Goodness, Reframed*.[187] We believe these three concepts should also be addressed in today's schools, embedded in the various subjects, but also having their own specialized emphasis.

In this technological century, it appears we are living in an age, as Gardner suggests, of "truthiness" and Twitter. Students are accessing much information from technology without assessing its accuracy, its truthfulness. Postmodernists suggest that belief in truth implies a modern rigidity. There are many truths within all realms: "political, economic, social, cultural"—and certainly educational.[188]

Gardner denotes that beauty describes the property of experiences, and we would add objects. Gardner indicates, "to be deemed beautiful, an experience must exhibit three characteristics: it must be interesting enough to behold, it must have a form that is memorable, and it must invite revisiting."[189] We seem blind to beauty, our eyes captured by our cell phones. Do we attempt to measure the beauty of the various disciplines, and should we?

Gardner describes goodness as relating to the interactions among humans.[190] We add that this quality also refers to our relations with the flora and fauna of the world and their and our environments. Yet we are not doing much in our schools to nurture goodness in any specific ways. We are not evaluating whether our students have this quality.

There is a need in this century for expanding what and how we evaluate our students, ourselves. All learning does not occur in schools; how to engage in self-evaluation is needed when functioning within the community. Perhaps the most basic questions in evaluation are not "What do you know?" or "What can you do?" but rather, "Who are you? What can you contribute to the world community?"

Conclusion

Evaluation addresses the value and effectiveness of curricular matters and activities. It centers on both teachers' and students' actions within the educational arena, primarily the classroom. Today, there is much debate regarding evaluation, primarily with demands that we must assess more effectively the actions of teachers and the learnings of students. There are clarion calls for teachers to be more effective in their pedagogical approaches and for students to achieve more and to attain higher standards to be competitive in the world community. These calls exist under the twin banners of *standards* and *accountability*.

Much talk about evaluation and, particularly, testing reveals a "buy in" by many people that education is a "business within the marketplace" and that its effectiveness should be judged with the same metric by which we judge workers and businesses. Productivity, attaining business goals, meeting quotas, and meeting market expectations are all ways to determine whether a business is meeting what it has set out to do. Schools should do the same.

This argument essentially reflects a scientific, modernist approach to evaluation. However, educators primarily in the humanistic, postmodernist camp of evaluation counter that schools are not making cars, processing mortgages, raising corn, or producing televisions or other electronics. You can count cars produced in a certain time period and make a judgment as to efficiency of production. Not so, many educators argue, with students' learnings. Certainly, you can compare test scores, and this seems to be the major metric for determining the effectiveness of teachers and the amount of student learnings. However, many involved in evaluation debate this query: What do test scores really say other than someone attained a 95 percent or is at the ninth stanine, and someone else got an 85 percent and is at the eighth stanine? And what do such comparisons really mean?

The current dialogue does indicate that evaluation addresses complex activities within complex contexts. There are myriad voices within these contexts, all driven by particular agendas. It behooves us to be knowledgeable about the clusters of procedures that deal with people as well as programs. Much dialogue regarding evaluation seems to exist within clouds of fear, confusion, ignorance, myopic thinking, and of course, enlightened ruminations. These dialogues involve individuals and groups of all stripes: educational, social, business, political, and even religious. Within these stripes we have stratifications of views, beliefs, aspirations, and attitudes. And within the stratifications we have degrees of certainties, uncertainties, stubbornness, and tolerance. This being the current state of affairs regarding educational evaluation, we should be mindful that evaluation not only assesses learning, but also promotes and nourishes it.

Discussion Questions

1. What are the nature and purpose of evaluation?
2. How do scientific, modernist and humanistic, postmodernist approaches differ in their assumptions?
3. How do formative and summative evaluation differ?
4. Describe the evaluation models recommended by Elliot Eisner.
5. What are the various issues that may be faced by minorities with regard to meritocratic education?

Notes

1. Peter Taubman, *Teaching by Numbers* (New York: Routledge, 2009), p. 12.
2. Ibid.
3. David C. Berliner and Gene V. Glass, *50 Myths & Lies that Threaten America's Public Schools* (New York: Teachers College Press, 2014), p. 11.
4. Ibid.
5. E. P. Cubberley, *Public School Administration* (Boston: Houghton Mifflin, 1916), p. 338, cited in Wayne Au, *Unequal by Design: High-Stakes Testing and the Standardization of Inequality* (New York: Routledge, Taylor & Francis Group, 2009), p. 19.
6. *Educational Leadership, STEM for All* (Alexandria, VA: ASCD, December 2014–January 2015).
7. Berliner and Glass, *50 Myths & Lies that Threaten America's Public Schools*, p. 14.
8. Taubman, *Teaching by Numbers*.
9. James W. Pellegrino, Naomi Chudowsky, and Robert Glaser, eds., *Knowing What Students Know: The Science and Design of Educational Assessment* (Washington, DC: National Academy Press, 2001).
10. Taubman, *Teaching by Numbers*, p. 29.
11. Pellegrino, Chudowsky, and Glaser, *Knowing What Students Know: The Science and Design of Educational Assessment*.

12. Maxine Greene, *Releasing the Imagination: Essays on Education, the Arts, and Social Change* (San Francisco: Jossey-Bass, 1995).
13. Pellegrino, Chudowsky, and Glaser, *Knowing What Students Know: The Science and Design of Educational Assessment.*
14. Au, *Unequal by Design,* p. 49.
15. Gary W. Ritter and Nathan C. Jensen, "The Delicate Task of Developing an Attractive Merit Pay Plan for Teachers," *Phi Delta Kappan* (May 2010), pp. 32–37.
16. Chris S. Hulleman and Kenneth E. Barron, "Performance Pay and Teacher Motivation: Separating Myth from Reality," *Phi Delta Kappan* (May 2010), pp. 27–31.
17. Ibid.
18. Matthew C. Springer and Catherine P. Gardner, "Teacher Pay for Performance: Context, Status, and Direction," *Phi Delta Kappan* (May 2010), pp. 8–15.
19. Ibid.
20. Berliner and Glass, *50 Myths & Lies that Threaten America's Public Schools,* p. 59.
21. Ibid., p. 61.
22. Pellegrino, Chudowsky, and Glaser, *Knowing What Students Know: The Science and Design of Educational Assessment.*
23. Ibid., p. 43.
24. Ibid.
25. Ibid.
26. David E. Tanner, *Assessing Academic Achievement* (Boston: Allyn & Bacon, 2001).
27. Catherine S. Taylor and Susan Bobbitt Nolen, *Classroom Assessment,* 2nd ed. (Upper Saddle River, NJ: Pearson, 2008).
28. Lisa Carter, *Total Instructional Alignment: From Standards to Student Success* (Bloomington, IN: Solution Tree Press, 2007).
29. Harriet Talmage, "Evaluating the Curriculum: What, Why and How," *National Association for Secondary School Principals* (May 1985), pp. 1–8.
30. Taylor and Nolen, *Classroom Assessment.*
31. Michael Fullan, ed., *The Challenge of Change,* 2nd ed. (Thousand Oaks, CA: Corwin, 2009), p. 25.
32. Ibid.
33. L. Lezotte and K. McKee, *Assembly Required: A Continuous School Improvement System* (Okemos, MI: Effective Schools Product, LTD, 2002), cited in Carter, *Total Instructional Alignment: From Standards to Student Success,* p. 55.
34. Blaine R. Worthen and James R. Sanders, *Educational Evaluation: Alternative Approaches and Practical Guidelines,* 2nd ed. (New York: Longman, 1987), pp. 22–23.
35. Abbie Brown and Timothy D. Green, *The Essentials of Instructional Design* (Upper Saddle River, NJ: Pearson, 2006).
36. Wilhelmina Savenye, "Evaluating Web-Based Learning Systems and Software," in Norbert M. Seel and Sanne Dijkstra, eds., *Curriculum, Plans, and Processes in Instructional Design: International Perspectives* (Mahwah, NJ: Lawrence Erlbaum Associates, 2004), pp. 309–330.
37. Daniel L. Stufflebeam, *Educational Evaluation and Decision Making* (Itasca, IL: Peacock, 1971), p. 25.
38. Collin J. Marsh and George Willis, *Curriculum: Alternative Approaches, Ongoing Issues,* 4th ed. (Upper Saddle River, NJ: Pearson, 2007), p. 266.
39. Kenneth A. Sirotnik and Jeannie Oakes, "Evaluation as Critical Inquiry: School Improvement as a Case in Point," in K. A. Sirotnik, ed., *Evaluation and Social Justice: Issues in Public Education* (San Francisco: Jossey-Bass, 1990), pp. 37–60.
40. *Merriam-Webster's Collegiate Dictionary,* 11th ed. (Springfield, MA: Merriam-Webster, 2004), p. 582.
41. Donald Blumenfeld-Jones, "Dance Curricula Then and Now: A Critical Historical-Hermeneutic Evaluation," in William M. Reynolds and Julie A. Webber, *Expanding Curriculum Theory: Dis/Positions and Lines of Flight* (Mahwah, NJ: Lawrence Erlbaum Associates, 2004), pp. 125–153.
42. Fred N. Kerlinger, *Behavioral Research: A Conceptual Approach* (New York: Holt, Rinehart and Winston, 1979).
43. Brown and Green, *The Essentials of Instructional Design.*
44. Michael Scriven, "The Methodology of Evaluation," in J. R. Gress and D. E. Purpel, eds., *Curriculum: An Introduction to the Field,* 2nd ed. (Berkeley, CA: McCutchan, 1988), pp. 340–412; and Blaine R. Worthen and Vicki Spandel, "Putting the Standardized Test Debate in Perspective," *Educational Leadership* (February 1991), pp. 65–69.
45. Patrick Slattery, *Curriculum Development in the Postmodern Era: Teaching and Learning in an Age of Accountability* (New York: Routledge, Taylor & Francis Group, 2013), p. 127.
46. Ibid., p. 119.
47. Ibid.
48. Ibid.
49. Ibid., p. 127.
50. Ibid., p. 119.
51. William E. Doll Jr., "Post-Modernism's Utopian Vision," in Donna Trueit, ed., *Pragmatism, Post-Modernism, and Complexity Theory: The "Fascinating Imaginative Realm" of William E. Doll, Jr.* (New York: Routledge, Taylor & Francis Group, 2012), pp. 144–152.
52. Ibid.
53. Ibid.
54. Ibid., p. 148.
55. Ibid., p. 149.
56. Ibid., p. 152.
57. Richard L. Curwin, "Can Assessments Motivate?" *Educational Leadership* (September 2014), pp. 38–40.
58. Ibid., p. 38.
59. Savenye, "Evaluating Web-Based Learning Systems and Software."
60. Taylor and Nolen, *Classroom Assessment.*
61. Lee J. Cronbach, *Designing Evaluations of Educational and Social Programs* (San Francisco: Jossey-Bass, 1982), p. 24.

62. Taylor and Nolen, *Classroom Assessment.*
63. Ibid.
64. Ibid.
65. Gina Schuyler Ikemoto and Julie A. Marsh, "Cutting through the 'Data-Driven' Mantra: Different Conceptions of Data-Driven Decision Making," in Pamela A. Moss, ed., *Evidence and Decision Making,* 106th Yearbook of the National Society for the Study of Education, Part 1 (Malden, MA: Distributed by Blackwell Publishing, 2007), pp. 105–131.
66. Ibid., p. 111.
67. James P. Comer, *What I Learned in School* (San Francisco: Jossey-Bass, 2009), p. 137.
68. Ibid.
69. Doll, "Post-Modernism's Utopian Vision," p. 145.
70. William A. Firestone and Raymond A. Gonzalez, "Culture and Processes Affecting Data Use in School," in Moss, *Evidence and Decision Making,* pp. 132–154.
71. Ibid., p. 141.
72. Ibid., p. 49.
73. Ibid.
74. Ibid.
75. Greene, *Releasing the Imagination, Essays on Education, the Arts, and Social Change.*
76. George F. Madaus and Thomas Kellaghan, "Curriculum Evaluation and Assessment," in Philip W. Jackson, ed., *Handbook of Research on Curriculum* (New York: Macmillan, 1992), pp. 119–154.
77. Ibid.
78. Pepi Leistyna, Arlie Woodrum, and Stephen A. Sherblom, *Breaking Free: The Transformative Power of Critical Pedagogy* (Cambridge, MA: Harvard Educational Review, 1999).
79. Ernest R. House, "Assumptions Underlying Evaluation Models," in G. F. Madaus, ed., *Evaluation Models: Viewpoints on Educational and Human Services* (Hingham, MA: Kluwer, 1983), pp. 45–64.
80. Worthen and Sanders, *Educational Evaluation: Alternative Approaches and Practical Guidelines.*
81. Scriven, "The Methodology of Evaluation."
82. Savenye, "Evaluating Web-Based Learning Systems and Software."
83. Brown and Green, *The Essentials of Instructional Design.*
84. Brent Duckor, "Formative Assessment in Seven Good Moves," *Educational Leadership* (March 2014), pp. 28–29.
85. Ibid., pp. 28–32.
86. Frederick Erickson, "Some Thoughts on 'Proximal' Formative Assessment in Student Learning," in Moss, *Evidence and Decision Making,* pp. 186–216.
87. W. James Popham, *Transformative Assessment* (Alexandria, VA: ASCD, 2008).
88. Ibid.
89. Allan Collins and Richard Halverson, *Rethinking Education in the Age of Technology* (New York: Teachers College Press, 2009).
90. Mike Schmoker, *Results Now* (Alexandria, VA: ASCD, 2006), pp. 130–131.
91. Taylor and Nolen, *Classroom Assessment.*
92. Savenye, "Evaluating Web-Based Learning Systems and Software."
93. Doll, "Post-Modernism's Utopian Vision."
94. Ibid.
95. D. L. Kirkpatrick, *Evaluating Training Programs: The Four Levels* (San Francisco: Berrett-Koehler, 1994), cited in Brown and Green, *The Essentials of Instructional Design.*
96. Ibid., pp. 249–250.
97. Ibid., p. 250.
98. Erickson, "Some Thoughts on 'Proximal' Formative Assessment in Student Learning," p. 190.
99. Ibid., p. 191.
100. Ibid.
101. Taubman, *Teaching by Numbers.*
102. Ibid.
103. Ibid.
104. Robert L. Thorndike, *Applied Psychometrics* (Boston: Houghton Mifflin, 1982).
105. H. H. Giles, S. P. McCutchen, and A. N. Zechiel, *Exploring the Curriculum* (New York: Harper & Row, 1942); and R. E. Smith and Ralph W. Tyler, *Appraising and Recording Student Progress* (New York: Harper & Row, 1942).
106. Robert E. Stake, "The Countenance of Educational Evaluation," *Teachers College Record* (April 1967), pp. 523–540.
107. Stufflebeam, *Educational Evaluation and Decision Making.*
108. Ibid., p. 229.
109. Taubman, *Teaching by Numbers.*
110. Ibid.
111. Sirotnik and Oakes, "Evaluation as Critical Inquiry: School Improvement as a Case in Point."
112. Taubman, *Teaching by Numbers.*
113. Slattery, *Curriculum Development in the Postmodern Era: Teaching and Learning in an Age of Accountability,* p. 119.
114. Ibid.
115. Taubman, *Teaching by Numbers.*
116. J. F. Lyotard, *The Postmodern Condition: A Report on Knowledge* (Minneapolis: University of Minnesota Press, 1989), cited in Taubman, *Teaching by Numbers.*
117. Elliot W. Eisner, *The Enlightened Eye* (Upper Saddle River, NJ: Merrill, 1998).
118. Ibid.
119. Ibid., p. 80.
120. Ibid.
121. Slattery, *Curriculum Development in the Postmodern Era: Teaching and Learning in an Age of Accountability,* p. 247.
122. M. Parlett and D. Hamilton, "Evaluation as Illumination: A New Approach to the Study of Innovative Programs,"

in G. V. Glass, ed., *Evaluation Studies Review Annual* (Beverly Hills, CA: Sage, 1976).
123. Greene, *Releasing the Imagination: Essays on Education, the Arts, and Social Change.*
124. Parker J. Palmer, *The Courage to Teach: Exploring the Inner Landscape of a Teacher's Life* (San Francisco: Jossey-Bass, 1998).
125. Charles I. Parker, "Preceptor and Pupil," *Daily Inter-Ocean* (Chicago: January 11, 1878), quoted in William J. Reese, *Testing Wars in the Public Schools: A Forgotten History* (Cambridge, MA: Harvard University Press, 2013), p. 1.
126. L. E. Rector, comments from educator for Jersey City, New Jersey, 1895, quoted in Reese, *Testing Wars in the Public Schools: A Forgotten History,* p. 222.
127. Reese, *Testing Wars in the Public Schools: A Forgotten History,* p. 231.
128. Ibid., p. 232.
129. Ibid.
130. A. Ronnell, *The Test* (Urbana, IL: University of Illinois Press, 2005), cited in Taubman, *Teaching by Numbers,* p. 17.
131. Berliner and Glass, *50 Myths & Lies that Threaten America's Public Schools,* p. 12.
132. Ibid., p. 11.
133. Au, *Unequal by Design,* pp, 122–123.
134. Taylor and Nolen, *Classroom Assessment,* p. 203.
135. Wayne Au, "High-Stakes Testing and Curriculum Control: A Qualitative Metasynthesis," pp. 235–251, in David J. Flinders and Stephen J. Thornton, eds., *The Curriculum Studies Reader,* 4th ed. (New York: Routledge, Taylor & Francis Group, 2013), p. 236.
136. G. Orfield and J. Wald, "Testing, Testing: The High-Stakes Testing Mania Hurts Poor and Minority Students the Most," *Nation* (2000), cited in Au, "High-Stakes Testing and Curriculum Control: A Qualitative Metasynthesis," p. 38.
137. Au, *Unequal by Design,* p. 87.
138. Ibid., p. 88.
139. Ibid., p. 89.
140. Taubman, *Teaching by Numbers.*
141. Brown and Green, *The Essentials of Instructional Design.*
142. Springer and Gardner, "Teacher Pay for Performance: Context, Status, and Direction."
143. Ibid.
144. G. Bowker, "Time, Money, and Biodiversity," in A. Ong and S. Collier, eds., *Global Assemblages: Technology, Politics and Ethics as Anthropological Problems* (Malden, MA: Blackwell, 2005), p. 109, cited in Taubman, *Teaching by Numbers,* p. 117.
145. Taubman, *Teaching by Numbers.*
146. Taylor and Nolen, *Classroom Assessment.*
147. Alfie Kohn, *The Schools Our Children Deserve* (Boston: Houghton Mifflin Company, 1999).
148. Ibid.
149. Taylor and Nolen, *Classroom Assessment.*
150. Ibid.
151. Marsh and Willis, *Curriculum: Alternative Approaches, Ongoing Issues.*
152. W. James Popham, "A Test Is a Test Is a Test—Not!" *Educational Leadership* (December 2006–January 2007), pp. 88–89.
153. Ibid., p. 88.
154. Ibid.
155. Taylor and Nolen, *Classroom Assessment.*
156. Marsh and Willis, *Curriculum: Alternative Approaches, Ongoing Issues.*
157. Taylor and Nolen, *Classroom Assessment.*
158. Tanner, *Assessing Academic Achievement.*
159. W. James Popham, "Criterion-Referenced Measurement: Half a Century Wasted?" *Educational Leadership* (March 2014), p. 65.
160. Ibid., pp. 64–65.
161. Ibid.
162. Allan C. Ornstein, "Comparing and Constructing Norm-Referenced and Criterion-Referenced Tests," *NASSP Bulletin* (1993).
163. Brown and Green, *The Essentials of Instructional Design.*
164. Pellegrino, Chudowsky, and Glaser, *Knowing What Students Know: The Science and Design of Educational Assessment.*
165. Carol A. Meyer, "What's the Difference between 'Authentic' and 'Performance' Assessment?" *Educational Leadership* (May 1992), pp. 39–40.
166. Ibid.
167. Bruce Frazee and Rose Ann Rudnitski, *Integrated Teaching Methods* (Albany, NY: Delmar, 1995).
168. Dennie Palmer Wolf and Sean F. Reardon, "Access to Excellence through New Forms of Student Assessment," in Joan Boykoff Baron and Dennie Palmer Wolf, eds., *Performance-Based Student Assessment: Challenges and Possibilities,* Ninety-fifth Yearbook of the National Society for the Study of Education, Part 1 (Chicago: University of Chicago Press, 1996).
169. Linda Darling-Hammond and Jacqueline Ancess, "Authentic Assessment and School Development," in Baron and Wolf, *Performance-based Student Assessment: Challenges and Possibilities.*
170. George E. Hein, "A Progressive Education Perspective on Evaluation," in Brenda S. Engel with Anne C. Martin, *Holding Values: What We Mean by Progressive Education* (Portsmouth, NH: Heinemann, 2005), pp. 176–185.
171. Taylor and Nolen, *Classroom Assessment.*
172. Ibid.
173. Ibid.
174. Ibid.
175. Au, *Unequal by Design.*
176. Ibid.
177. Ibid.
178. N. Lemann, *The Big Test: The Secret History of the American Meritocracy* (New York: Farrar, Straus, and Giroux, 1999); and P. Sacks, *Standardized Minds: The*

High Price of America's Testing Culture and What We Can Do to Change It (Cambridge, MA: Perseus Books, 1999), cited in Au, *Unequal by Design,* pp. 45–46.

179. David P. Ericson, "Social Justice, Evaluation and the Educational System," in Sirotnik, *Evaluation and Social Justice: Issues in Public Education,* pp. 5–22.
180. Pellegrino, Chudowsky, and Glaser, *Knowing What Students Know: The Science and Design of Educational Assessment.*
181. Deborah Meier, *In Schools We Trust* (Boston: Beacon Press, 2002), p. 109.
182. Pellegrino, Chudowsky, and Glaser, *Knowing What Students Know: The Science and Design of Educational Assessment.*
183. Ibid.
184. Deborah Landry, "Teachers' (K–5) Perceptions of Student Behaviors during Standardized Testing," in Barbara Slater Stern, ed., *Curriculum and Teaching Dialogue* (Greenwich, CT: Information Age Publishing, 2006), pp. 29–40.
185. Peg Tyre, "The New First Grade: Too Much Too Soon," *Newsweek* (September 11, 2006), pp. 34–44.
186. Ibid., p. 36.
187. Howard Gardner, *Truth, Beauty, and Goodness Reframed* (New York: Basic Books, 2011).
188. Ibid., p. 192.
189. Ibid., p. xi.
190. Ibid.