

# Testing Hazards in Performance Contracting<sup>1</sup>

Robert E. Stake<sup>2</sup>

In the first federally sponsored example of performance contracting for the public schools, Dorsett Educational Systems of Norman, Oklahoma, contracted to teach reading, mathematics, and study skills to over 200 poor-performance junior and senior high school students in Texarkana. Commercially available, standardized, general-achievement tests were used to measure performance gains.

Are such tests suitable for measuring specific learnings? To the person little acquainted with educational testing, it appears that performance testing is what educational tests are for. The testing specialist knows better. General achievement tests have been developed to measure correlates of learning, not learning itself.

Such tests are indirect measures of educational gains. They provide correlates of achievement rather than direct evidence of achievement. Correlation of these test scores with general learning is often high, but such scores correlate only moderately with performance on many specific education tasks. Tests can be built to measure specific competence, but there is relatively little demand for them. Many of those tests (often called criterion referenced tests) do a poor job of predicting later performance of either a specific or a general nature. General achievement tests predict better. The test developer's basis for improving tests has been to work toward better prediction of later performance rather than better measurement of present performance. Assessment of what a student is now capable of doing is not the purpose of most standardized tests. Errors and hazards abound, especially when these general achievement tests are used for performance contracting. Many of the hazards remain even with the use of criterion-referenced tests or any other performance observation procedures.

One of the hazards in performance contracting is that many high-priority educational objectives for various reasons and in various ways will be cast aside while massive attention is given to other high-priority objectives. This hazard is not unrelated to testing but will not be discussed here. This article will identify the major obstacles to gathering direct evidence of performance gain on targeted objectives.

## Errors of Testing

Answering a *National School Board Journal* (November, 1970) questionnaire on performance contracting, a New Jersey board member said:

Objectives must be stated in simple, understandable terms. No jargon will do and no subjective goals can be tolerated. Neither can the nonsense about there being some mystique that prohibits objective measurement of the educational endeavor.

Would that our problems withered before stern resolve. But neither wishing nor blustering rids educational testing of its errors.

---

<sup>1</sup> Published in *Phi Delta Kappan*, June, 1971, pages 583-588.

<sup>2</sup> Robert E. Stake is associate director of the Center for Instructional Research and Curriculum Evaluation, College of Education, University of Illinois.

Just as the population census and the bathroom scales have their errors, educational tests have theirs. The technology and theory of testing are highly sophisticated; the sources of error are well known.<sup>3</sup> Looking into the psychometrist's meaning of a theory of testing, one finds a consideration of ways to analyze and label the inaccuracies in test scores. There is a mystique, but there is also simple fact: No one can eliminate test errors. Unfortunately, some errors are large enough to cause wrong decisions about individual children or school district policy.

Some educators and social critics consider the whole idea of educational testing to be a mistake.<sup>4</sup> Unfortunate social consequences of testing, such as the perpetuation of racial discrimination<sup>5</sup> and pressures to cheat,<sup>6</sup> continue to be discussed. But, as expected, most test specialists believe that the promise in testing outweighs these perils. They refuse responsibility for gross misuse of their instruments and findings and concentrate on reducing the errors in tests and test programs.

Some technical errors in test scores are small and tolerable. But some testing errors are intolerably large. Today's tests can, for example, measure vocabulary word-recognition skills with sufficient accuracy. They cannot, however, adequately measure listening comprehension or the ability to analyze the opposing sides of an argument.

Contemporary test technology is not refined enough to meet all the demands. In performance contracting the first demand is for assessment of performance. Tests do their job well when the performance is highly specific--when, for example, the student is to add two numbers, recognize a misspelled word, or identify the parts of a hydraulic lift. When a teacher wants to measure performances that require more demanding mental processes, such as conceptualizing a writing principle or synthesizing a political argument, performance tests give us less dependable scores.<sup>7</sup>

**Unreached potentials.** Many educators believe that the most human of human gifts - the emotions, the higher thought processes, interpersonal sensitivity, moral sense--are beyond the reach of psychometric testing. Most test specialists disagree. While recognizing an ever-present error component, they believe that anything can be measured. The credo was framed by E. L. Thorndike in 1918: "Whatever exists at all exists in some amount." Testing men believe it still. They are not so naive as to think that any human gift will manifest itself in a 45-minute paper-and-pencil test. They do believe that, given ample opportunity to activate and observe the examinee, any trait, talent, or learning that manifests itself in behavior can be measured with reasonable accuracy. The total cost of measuring may be 100 times that of administering the usual tests, but they believe it can be done. The final observations may rely on professional judgment, but this could be reliable and validated judgment. A question for most test specialists, then, is not "Can complex educational outcomes be measured?" but "Can complex educational outcomes be measured with the time and personnel and facilities available?"

---

<sup>3</sup> Frederick M. Lord and Melvin R. Novick, *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley, 1968.

<sup>4</sup> Banesh Hoffman, *The Tyranny of Testing*. New York: Collier Books, 1962; and Theodore R. Sizer, "Social Change and the Uses of Educational Testing: An Historical View," paper presented at the *Invitational Conference on Testing Problems*, New York, October, 1970.

<sup>5</sup> David A. Goslin, "Ethical and Legal Aspects of the Collection and Use of Educational Information," paper presented at the *Invitational Conference on Testing Problems*, New York, October, 1970.

<sup>6</sup> Barry R. McGhan, "Accountability as a Negative Reinforcer," *American Teacher*, November, 1970, p. 13.

<sup>7</sup> Benjamin S. Bloom et al., *A Taxonomy of Educational Objectives: Handbook I, the Cognitive Domain*. New York: McKay, 1956

When it is most important to know whether or not a child is reading at age-level, we call in a reading specialist, who observes his reading habits. She might test him with word recognition, syntactic decoding, and paragraph comprehension exercises. She would retest where evidence was inconclusive. She would talk to his teachers and his parents. She would arrive at a clinical description--which might be reducible to a statement such as "Yes, Johnny is reading at or above age-level." The scores we get from group reading tests can be considered estimates of such an expert judgment. These objective test scores correlate positively with the more valid expert judgments. Such estimates are not direct measurements of what teachers or laymen mean by "ability to read," nor are they suitably accurate for diagnostic purposes. Achievement gains for a sizable number of students will be poorly estimated. It is possible that the errors in group testing are so extensive that--when fully know--businessmen and educators will refuse to accept them as bases for contract reimbursement.

**Professional awareness.** Classroom teachers and school principals have tolerated standardized test errors because they have not been obligated to make crucial decisions on the basis of test scores. Actually, in day-to-day practice they seldom use test scores.<sup>8</sup> When they do, they combine them with other knowledge to estimate a child's progress in school and to guide him into an appropriate learning experience. They do not use tests as a basis for assessing the quality of their own teaching.

In performance contracting, the situation is drastically changed; tests are honored as the sole basis for contract reimbursement. The district will pay the contractor for performance improvement. An error in testing means money misspent. Course completion and reimbursement decisions are to be made without reliance on the knowledge and judgment of a professional observer, without asking persons who are closest to the learning (the teacher, the contractor, and the student) whether or not they *see* evidence of learning. Decisions are to be made entirely by objective and independent testing. Numerous human errors and technical misrepresentations will occur.

### Which Test Items?

It is often unrealistic to expect a project director to either find or create paper-and-pencil test items, administrable in an hour to large numbers of students by persons untrained in psychometric observation and standardized diagnostics, objectively scorable, valid for purposes of the performance contract, and readily interpretable. The more complex the training, the more unrealistic the expectation. One compromise is to substitute criterion test items measuring simple behaviors for those measuring the complex behaviors targeted by the training. For example, the director may substitute vocabulary-recognition test, items for reading-comprehension items or knowledge of components for the actual dismantling of an engine. The substitution may be reasonable, but the criterion test should be, validated against performances directly indicated by the objectives. It almost never has been. Without the validation the educator should be skeptical about what the test measures.

---

<sup>8</sup> J. Thomas Hastings, Philip. J. Runkel, and Dora E. Damrin, *Effects on Use of Tests by Teachers Trained in a Summer Institute*, Cooperative Research Project No. 702. Urbana, Ill.: Bureau of Educational Research, College of Education, University of Illinois, 1961.

It always is unrealistic to expect that the payoff from instruction will be apparent in the performances of learners at test-taking time.<sup>9</sup> Most tests evoke relatively simple behavior. Ebel wrote:

...most achievement tests...consist primarily of items testing specific elements of knowledge, facts, ideas, explanations, meanings, processes, procedures, relations, consequences, and so on.<sup>10</sup>

He went on to point out that more than simple recall is involved in answering even the simplest vocabulary item.

Much more complex behavior is needed for answering a reading comprehension item. These items clearly call for more than the literal meanings of the words read. The student must paraphrase and interpret - what we expect readers to be able to do.

These items and ones for problem solving and the higher mental processes do measure high-priority school goals - but growth in such areas is relatively slow. Most contractors will not risk basing reimbursement on the small chance that evidence of growth will be revealed by *these* criterion tests. Some of the complex objectives of instruction will be under-emphasized in the typical performance-contract testing plan.

The success of Texarkana's first performance - contract year is still being debated. Late winter (1969-70) test results looked good, but spring test results were disappointing.<sup>11</sup> Relatively simple performance items had been used. But the "debate" did not get into that. It started when the project's "outside evaluator" ruled that there had been direct coaching on most, if not all, of the criterion test items, which were known by the contractor during the school year. Critics claimed unethical "teaching for the test." The contractor claimed that both teaching and testing had been directed toward the same specific goals, as should be the case in a good performance contract. The issue is not only test choice and ethics; it includes the ultimate purpose of teaching.

**Teaching for the test.** Educators recognize an important difference between preparation for testing and direct coaching for a test. To prepare an examinee, the teacher teaches within the designated knowledge-skill domain and has the examinee practice good test-taking behavior (for example, don't spend too much time on overly difficult items; guess when you have an inkling though not full knowledge; organize your answer before writing an essay) so that relevant knowledge-skill is not obscured. Direct coaching teaches the examinees how to make correct responses to specific items on the criterion test.

This is an important difference when test items cover only a small sample of the universe of what has been taught or when test scores are correlates, rather than direct measurements, of criterion behavior. It ceases to be important when the test is set up to measure directly and

---

<sup>9</sup> Harry S. Broady, "Can Research Escape the Dogma of Behavioral Objectives?," *School Review*, November, 1970, pp. 43-56

<sup>10</sup> Robert L. Ebel, "When Information Becomes Knowledge," *Science*, January, 1971, pp. 130-31.

<sup>11</sup> Dean C. Andrew and Lawrence H. Roberts, "Final Evaluation Report on the Texarkana Dropout Prevention Program, Magnolia, Arkansas: Region VIII." Education Service Center, July 20, 1970 (mimeo). Commentaries on this report include Henry S. Dyer, "Performance Contracting: Too Simple a Solution for Difficult Problems," *The United Teacher*, November 29, 1970, pp. 1922; and Roger T. Lennon, "Accountability and Performance Contracting," paper presented at the annual meeting of the American Educational Research Association of the New York, February, 1971.

thoroughly that which has been taught. In this case, teaching for the test is exactly what is wanted.

Joselyn pointed out that the performance contractor and the school should agree in advance on the criterion procedure, though not necessarily on the specific items.<sup>12</sup> To be fair to the contractor, the testing needs to be reasonably close to the teaching. To be fair to parents, the testing needs to be representative of the domain of abilities *they* are concerned about. A contract to develop reading skills would not be satisfied adequately by gains on a vocabulary test, according to the expectations of most teachers. All parties need to know how similar the testing will be to the actual teaching.

**A dissimilarity scale.** Unfortunately, as Anderson observed,<sup>13</sup> the test specialist has not developed scales for describing the similarity between teaching and testing. This is a grievous failing. Educators have no good way to indicate how closely the tests match the instruction.

There are many ways for criterion questions to be dissimilar. They can depart from the information taught by: 1) syntactic transformation; 2) semantic transformation; 3) change in content or medium; 4) application, considering the particular instance; 5) inference, generalizing from learned instances; and 6) implication, adding last-taught information to generally known information. For examples of some of these transformations, see Table 1. Hively, Patterson, and Page;<sup>14</sup> Bormuth;<sup>15</sup> and Jackson<sup>16</sup> have discussed procedures for using some of these transformations to generate test items.

For any student the appropriateness of these items depends on prior and subsequent learning as well as on the thoroughness of teaching. Which items are appropriate will have to be decided at the scene. The least and most dissimilar items might be quite different in their appropriateness. The reading-comprehension items of any standardized achievement battery are likely to be more dissimilar to the teaching of reading than any of the "dissimilarities" shown in Table 1. Immediate instruction is not properly evaluated by highly dissimilar items, nor is scholarship properly evaluated by highly similar items. Even within the confines of performance contracting, both evaluations are needed.

For the evaluation of instruction, a large number of test items are needed for each objective that--in the opinion of the teachers--directly measure increase in skill or understanding. Items from standardized tests, if used, would be included item by item. For each objective, the item pool would cover all aspects of the objective. A separate sample of items would be drawn for the pretest and posttest for each student, and instructional success would be based on the collective gain of all students.

Creating such a pool of relevant, psychometrically sound test items is a major--but necessary--undertaking.<sup>17</sup> It is a partial safeguard against teaching for the test and against the use of inappropriate criteria to evaluate the success of instruction.

---

<sup>12</sup> E. Gary Joselyn, "Performance Contracting: What It's All About," paper presented at the Truth and Soul in Teaching Conference of the American Federation of Teachers, Chicago, January, 1971.

<sup>13</sup> Richard C. Anderson, "Comments on Professor Gagne's Paper," in *The Evaluation of Instruction*, ed. M. C. Wittrock and David E. Wiley. New York: Holt, Rinehart and Winston, 1970, pp. 126-33.

<sup>14</sup> Wells Hively II, Harry L. Patterson, and Sara H. Page, "A 'Universe-Defined' System of Arithmetic Achievement Tests," *Journal of Educational Measurement*, Winter, 1968, pp. 275-90.

<sup>15</sup> John Bormuth, *On the Theory of Achievement Test Items*. Chicago: University of Chicago Press, 1970.

<sup>16</sup> Rex Jackson, *Developing Criterion-Referenced Tests*. Princeton, N.J.: ERIC Clearing House on Tests, Measurement, and Evaluating, Educational Testing Service, June, 1970.

<sup>17</sup> Dorsett indicated the desirability of such an item pool in the original Texarkana proposal.

## What the Scores Mean

At first, performance contracting seemed almost a haven for the misinterpretation of scores. Contracts have ignored 1) the practice effect of pretesting,<sup>18</sup> 2) the origins of grade equivalents, 3) the "learning calendar," 4) the unreliability of gain scores, and 5) regression effects. Achievement may be spurious. Ignoring any one of these five is an invitation to misjudge the worth of the instruction.

---

**Table 1. An example of transformations of information taught into test questions.**

Information taught: Pt. Barrow is the northernmost town in Alaska.

Minimum transformation question: What is the northernmost town in Alaska?

Semantic syntactic transformation question: What distinction does Pt. Barrow have among Alaskan villages?

Context medium transformation question: The dots on the adjacent map represent Alaskan cities and towns. One represents Pt. Barrow. Which one?

Implication questions: What would be unusual about summer sunsets in Pt. Barrow?

---

**Grade-equivalent scores.** Standardized achievement tests have the appealing feature of yielding grade-equivalent scores. Each raw score, usually the number of items right, has been translated into a score indicating (for a student population forming a national reference group) the average grade placement of all students who got this raw score. These new scores are called "grade equivalents." Raw scores are not very meaningful to people unacquainted with the particular test; the grade equivalents are widely accepted by teachers and parents. Grade equivalents are common terminology in performance contracts. Unfortunately, grade equivalents are available from most publishers only for tests, not for test items. Thus the whole test needs to be used, in the way prescribed in its manual, if the grade equivalents are to mean what they are supposed to mean. One problem of using whole tests was discussed in the previous section. Another problem is that the average annual "growth" on most standardized tests is only a few raw-score points. Consider in Table 11 the difference between a grade equivalent of 5.0 and 6.0 within four of the most popular test batteries. Most teachers do not like to have their year's work summarized by so little change in performance. Schools writing performance contracts perhaps should be reluctant to sign contracts for which the distinction between success and failure is so small. But to do so requires the abandonment of grade equivalents.

**The learning calendar.** For most special instructional programs, criterion tests will be administered at the beginning of and immediately following instruction, often in the first and last weeks of school. A great deal of distraction occurs during those weeks, but other times for

---

<sup>18</sup> Not discussed here because of space limitations.

pretesting and posttesting have their hazards, too. Recording progress every few weeks during the year is psychometrically preferred, but most teachers are opposed to "all that testing."

---

**Table 11. Gain in items right needed to advance one grade equivalent on four typical achievement tests.**

	<i>Grade equivalent</i>		<i>Needed for an improvement of one grade equivalent</i>
	<i>5.0</i>	<i>6.0</i>	
Comprehensive Test of Basic Skills, Level 3: Reading Comprehension	20	23	3 items
Metropolitan Achievement Test, Intermediate Form B: Spelling	24	31	7 items
Iowa Tests of Basic Skills, Test A1: Arithmetic Concepts	10	14	4 items
Stanford Achievement Test, Form Intermediate 11: Word Meaning	18	26	8 items

---

Children learn year-round, but the evidence of learning that gets inked on pupil records comes in irregular increments from season to season. Winter is the time of most rapid advancement, summer the least. Summer, in fact, is a period of setback for many youngsters. Beggs and Hieronymus found punctuation scores to spurt more than a year between October and April but to drop almost half a year between May and September.<sup>19</sup> Discussing their reading test, Gates and MacGinitie wrote:

. . . in most cases, scores will be higher at the end of one grade than at the beginning of the next. That is, there is typically some loss of reading skill during the summer, especially in the lower grades.<sup>20</sup>

The first month or two after students return to school in the fall is the time for getting things organized and restoring scholastic abilities lost during the summer. According to some records, spring instruction competes poorly with other attractions. Thus, the learning year is a lopsided year, a basis sometimes for miscalculation. Consider the results of testing shown in Table III.

The six-week averages in Table III are fictitious, but they represent test performance in many classrooms. The mean growth for the year appears to be 1.3 grade equivalents. No acknowledgement is made that standardized test results in early September were poorer than

---

<sup>19</sup> Donald L. Beggs and Albert N. Hieronymus, Uniformity of Growth in the Basic Skills Throughout the School Year and During the Summer, *Journal of Educational Measurement*, Summer, 1968, pp. 91-97.

<sup>20</sup> Arthur I. Gates and Walter H. MacGinitie, Technical Manual for the Gates-MacGinitie Reading Tests. New York: Teachers College Press, Columbia University, 1965, p. 5.

those for the previous spring. For this example, the previous May mean (not shown) was 5.2. The real gain, then, for the year is 1.1 grade equivalents rather than the apparent 1.3. It would be inappropriate to pay the contractor for a mean gain of 1.3.

Another possible overpayment on the contract can result by holding final testing early and extrapolating the previous per-week growth to the weeks or months that follow. In Texarkana, as in most schools, spring progress was not as good as winter's. If an accurate evaluation of contract instructional services is to be made, repeated testing, perhaps a month-by-month record of learning performances, needs to be considered.<sup>21</sup>

**Unreliable gain scores.** Most performance contracts pay off on an individual student basis. The contractor may be paid for each student who gains more than an otherwise expected amount. This practice is commendable in that it emphasizes the importance of each individual learner and makes the contract easier to understand, but it bases payment on a precarious mark: the gain score.

Just how unreliable is the performance-test gain score? For a typical standardized achievement test with two parallel forms, A and B, we might find the following characteristics reported in the test's technical manual:

Reliability of Test A = +.84.  
 Reliability of Test B = +.84.  
 Correlation of Test A with Test B = +.81.

Almost all standardized tests have reliability coefficients at this level. Using the standard formula,<sup>22</sup> one finds a disappointing level of reliability for the measurement of improvement:

Reliability of gain scores (A-B or B-A) = +.16.

The test manual indicates the raw score and grade-equivalent standard deviations. For one widely used test, they are 9.5 items and 2.7 years, respectively. Using these values we can calculate the errors to be expected. *On the average*, a student's raw score would be in error by 2.5 items, grade equivalent would be in error by 0.72 years, and grade-equivalent *gain score* would be in error by 1.01 years. The error is indeed large.

Consider what this means for the not unusual contract whereby the student is graduated from the program, and the contractor is paid for his instruction, on any occasion that his performance score rises above a set value. Suppose--with the figures above--the student exits when his improvement is one grade equivalent or more. Suppose also, to make this situation simpler, that there is no intervening training and that the student is not influenced by previous testing. Here are three ways of looking at the same situation:

Suppose that a contract student takes a different parallel form of the criterion test on three successive days immediately following the pretest. The chances are better than 50-50 that on *one*

---

<sup>21</sup> L. Wrightman and W. P. Gorth, CAM: The New Look in Classroom Testing, Trend, Spring, 1969, pp. 56-57. Project CAM is described as a model for a continuous (perhaps biweekly) monitoring and recording of classroom performance.

<sup>22</sup> Robert L. Thorndike and Elizabeth Hagen, Measurement and Evaluation in Psychology and Education, 3rd ed. New York: Wiley, 1969, p. 197.

of these tests the student will gain a year or more in performance and appear to be ready to graduate from the program.

Suppose that three students are tested with a parallel form immediately after the pretest. The chances are better than 50-50 that one of the three students entirely due to errors of measurement will gain a year or more and appear ready to graduate.

Suppose that 100 students are admitted to contract instruction and pretested. After a period of time involving no training, they are tested again, and the students gaining a year are graduated. After another period of time, another test and another graduation. After the fourth terminal testing, even though no instruction has occurred, the chances are better than 50-50 that thirds of the students will be graduated.

In other words, owing to unreliability, gain scores can appear to reflect learning that actually does not occur.

The unreliability will give an equal number of false impressions of deteriorating performance. These errors (false gains and false losses) will balance out for a large group of students. If penal ties for losses are paid by the contractor at the rate bonuses are paid for gains, the contractor will not be over paid. But according to the way contracts are being written, typified in the examples above, the error in the gain scores does not balance out; it works in favor of the contractor. Measurement errors could be capitalized upon by unscrupulous promoters. Appropriate checks against these errors are built into the better contracts. Errors in individual gain scores can be reduced by using longer tests. A better way to indicate true gain is to calculate the discrepancy between actual and expected final performances.<sup>23</sup> Expectations can be based on the group as a whole or on an outside control group. Another way is to write the contract on the basis of mean scores for the group of students.<sup>24</sup> Corrections for the unreliability of gain scores are possible, but they are not likely to be considered if the educators and contractors are statistically naive.

**Regression effects.** Probably the source of the greatest misinterpretation of the effects of remedial instruction is regression effects. Regression effects are easily overlooked but need not be; they are correctable. For any pretest score, the expected regression effect can be calculated. Regression effects make the poorest scorers look better the next time tested. Whether measurements are error-laden or error free, meaningful or meaningless, when there is differential change between one measurement occasion and another (when there is less-than-perfect correlation), the lowest original scorers will make the greatest gains and the highest original scorers will make the least. On the average, posttest scores will, relative to their corresponding pretest scores, lie in the direction of the mean. This is the regression effect. Lord discussed this universal phenomenon and various ways to correct for it.<sup>25</sup>

The demand for performance contracts has occurred where conventional instructional programs fail to develop--for a sizable number of students--minimum competence in basic skills. Given a distribution of skill test scores, the lowest-scoring students--the ones most needing

---

<sup>23</sup> Ledyard R. Tucker, Fred Damarin, and Samuel Messick, A Base-Free Measure of Change, Research Bulletin RB-65-16. Princeton, N.J.: Educational Testing Service, 1965. This is a discussion of change Scores that are independent of and dependent on the initial standing of the learning. A learning curve fitted to test scores could be used to counter the unreliability of individual scores.

<sup>24</sup> This would have the increased advantage of discouraging the contractor from giving preferential treatment within the project to students who are in a position to make high payoff gains.

<sup>25</sup> Frederic M. Lord, Elementary Models for Measuring Change, in Problems in Measuring Change, ed. Chester W. Harris Madison, Wis.: University of Wisconsin Press, 1963, pp. 21-38.

assistance--are identified. It is reasonable to suppose that under unchanged instructional programs they would drop even farther behind the high-scoring students. If a retest is given, however, after any period of instruction (conventional or special) or of no instruction, these students will no longer be the poorest performers. Some of them will be replaced by others who appear to be most in need of special instruction. Instruction is not the obvious influence here--regression is. The regression effect is not due to test unreliability, but it causes some of the same misinterpretations. The contract should read that instruction will be reimbursed when gain exceeds that attributable to regression effects. The preferred evaluation design would call for control group(s) of similar students to provide a good estimate of the progress the contract students would have made in the absence of the special instruction.

---

**Table 111. Learning calendar for a typical fifth-grade class.**

	Month									
	S	O	N	D	J	F	M	A	M	
Mean achievement score	5.0	5.3	5.6	5.9	6.2	6.3				

---

### **The Social Process**

The hazards of specific performance testing and performance contracting are more than curricular and psychometric. Social and humanistic challenges should be raised, too. The teacher has a special opportunity and obligation to observe the influence of testing on social behavior. Performance contracting has the unique ability to put the student in a position of administrative influence. He can make the instruction appear better or worse than it actually is by his performance on tests. Even if he is quite young, the student will know that his good work will benefit the contractor. Sooner or later he is going to know that, if he tests poorly at the beginning, he can benefit himself and the contractor through his later achievement. Bad performances are in his repertoire, and he may be more anxious to make the contractor look bad than to make himself look good. Or he may be under undue pressure to do well on the posttests. These are pupil-teacher interactions that should be watched carefully. More responsibility for school control possibly should accrue to students, but performance contracts seem a devious way to give it.

To motivate the student to learn and to make him want more contract instruction, many contractors use material or opportunity-to-play rewards. (Dorsett used such merchandise as transistor radios.) Other behavior modification strategies are common. The proponents of such strategies argue that, once behavior has been oriented to appropriate tasks, the students can gradually be shifted, from Extrinsic rewards to intrinsic. That they can be shifted is probably true; that it will happen without careful, deliberate work by the instructional staff is unlikely. It is not difficult to imagine a performance-contract situation in which the students become even less responsive to the rewards of conventional instruction.

In mid-1971, performance contracting appears to be popular with the current administration in Washington because it encourages private businesses to participate in a traditionally public responsibility. It is popular among some school administrators because it affords new access to federal funds, because it is a way to get new talent working on old problems, and because the administrator can easily blame the outside agency and the government if the contract instruction is unsuccessful. It is unpopular with the American Federation of Teachers because it reduces the control the union has over school operations, and it reduces the teacher's role as a chooser of what learning students need most. Performance contracting is popular among most instructional technologists because it is based on well-researched principles of teaching and because it enhances their role in school operations.

The accountability movement as a whole is likely to be a success or failure on such sociopolitical items. The measurement of the performance of performance contracting is an even more hazardous procedure than the measurement of student performances.

## **Summary**

Without yielding to the temptation to undercut new efforts to provide instruction, educators should continue to be apprehensive about evaluating teaching on the basis of performance testing alone. They should know how difficult it is to represent educational goals with statements of objectives and how costly it is to provide suitable criterion testing. They should know that the commonsense interpretation of these results is frequently wrong. Still, many members of the profession think that evaluation controls are extravagant and mystical.

Performance contracting has emerged because people inside and outside the schools are dissatisfied with the instruction some children are getting. Implicit in the contracts is the expectation that available tests can measure the newly promised learning. The standardized test alone cannot measure the specific outcomes of an individual student with sufficient precision.