DOCUMENT RESUME

ED 049 520 EA 003 358

AUTHOR Lennon, Roger T.
TITLE Accountability and Performance Contracting.
PUB DATE 5 Feb 71
NOTE 21p.; Speech presented at American Educational Research Association Annual Meeting. (55th, New York, New York, February 4-7, 1971)

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Academic Achievement, *Educational Accountability, Evaluation Techniques, Instructional Programs, *Measurement Instruments, Measurement Techniques, Models, *Performance Contracts, Skill Development, Speeches, *Test Reliability, *Test Validity, Textbook Standards

ABSTRACT

This report defines the concepts and some of the problems of accountability and performance contracting with special emphasis on measurement problems in the latter. Measurement problems involve both the validity and the reliability of standardized achievement tests as a basis for reimbursing a contractor. The author suggests the use of criterion referenced tests as a possible remedy to some of these problems, but cautions that results should be translatable into units that will yield measures of gain or growth. Related documents are EA 003 347, EA 003 356, EA 003 391, and EA 003 387. (JF)

ACCOUNTABILITY AND PERFORMANCE CONTRACTING*

Roger T. Lennon, Senior Vice President, Harcourt Brace Jovanovich, Inc.
and President, The Psychological Corporation

The perspective that I bring to this discussion of accountability and performance contracting is one growing out of my association with an organization which proffers instructional materials and services, measuring instruments, and evaluative services. The opinions that I shall voice are not necessarily those of my colleagues, nor do they in any sense constitute a declaration of official policy of the house. They are an outgrowth of protracted discussions in which my associates and I have attempted to define how we might, as responsible publishers, act in relation to requests for "guarantees" of the performance of our materials and services, and equally protracted discussions of the measurement and evaluation problems in performance-contract arrangements. We are in fact providing some instructional materials and support services on a contract basis, and we are furnishing measurement instruments for use in evaluating many performance contract programs; but our experience, like everyone else's at this stage, is limited. Having in mind the Dorsett experience in Texarkana, I am tempted to put it that, with respect to several performance-contract situations, we were the successful bidder – we did not get the contract.

ACCOUNTABILITY

Few terms that have come into the language of education have evoked the ready acceptance and nearly universal approbation that has attended the term "accountability." The reasons are not hard to find. Most observers credit Dr. Leon Lessinger with the earliest and most vigorous advocacy of both the concept and the term, during the time when he was serving as Deputy Commissioner in the Office of Education. In that role he witnessed the frustration felt by many members of Congress as they sought to assess the efficacy of the federal

*Invited address to the American Educational Research Association, February 5, 1971, Americana Hotel, New York City.

monies being expended on education, and as they sought to develop policy for educational expenditures. Legislators were distressed to learn how little could be asserted with confidence, for example, about the impact of funds expended under Title I of ESEA. Schooled in the "more bang for the buck" approach espoused in Department of Defense budgeting, they raised questions about the cost-effectiveness of various educational programs which the Office of Education and school people found very difficult to answer. Increasingly, their concern was echoed by school boards across the country confronted with never-ending requests for additional funds, and by taxpayers beginning to wonder whether the ever-increasing expenditures were really buying more or better education for their children. Dissatisfaction with the lack of success that attended most efforts to improve the level of achievement of inner-city and disadvantaged pupils, and the growing concern that schools be rendered more responsive to the communities which they served, particularly in the large metropolitan centers, combined to create a readiness for the proposition that school officials at every level should in some fashion be made accountable, that is, responsible for bringing about learning that could be shown to be commensurate with, or satisfactory in relation to, the resources being committed to the effort.

Educators, so the accountability message ran, habitually sought to justify requests for funds in terms of needs such as buildings, books in the libraries, books in the pupils' hands, teachers' salaries, learning equipment – in short, process variables - rather than in terms of manifest product - pupil learning of demonstrable magnitude. Some accountability spokesmen, to be sure, grossly overstated the case that school men had not been concerned with pupil achievement. We all know better. To pretend that only under the goad of accountability would we recognize that the effectiveness of education must be sought in evidences of pupil learning is to impugn needlessly the good sense and good will of countless

generations of educators. After all, the notion that compensation of an instructor should depend on student attainment goes back at least to the time of the medieval universities. We are told that at the University of Bologna in the 15th century, student-enacted statutes required that the "professor start his lectures at the beginning of the book, cover each section sequentially, and complete the book by the end of the term"; if the professor failed to achieve the schedule, he forfeited part of funds that he himself had had to deposit at the beginning of the term! The concern of governmental bodies that they were getting their educational dollar's worth was manifest in 1911 when the Board of Estimate of the City of New York, critical of the demands made by the Board of Education on the city's treasury, launched a comprehensive survey of the city's schools, one aspect of which was an analysis of the tested arithmetic achievement of its pupils. The first wave of textbooks in educational measurement — those published say between 1912 and 1922 or 1923 - abound with references to the utility of standardized achievement test results as indicators of the effectiveness of schools and even of teaching efficiency. So, the notion of pupil learning as the proper criterion in the establishment of accountability is in no sense new. Accountability, we might say, is an idea whose time has come — again, or perhaps an idea whose time is always.

I do not allude to the historical concern with student outcomes in any way to minimize or disparage the importance of the current concern with accountability—quite the contrary. I think that the perennial concern with student outcomes as indices of effectiveness attests both to the validity of the notion of accountability and to the extreme difficulties that have been experienced over the years in implementing the concept. What is new about accountability as currently advocated is the realization of the necessity for relating output in some sense to input, defining input as all professional staff effort,

financial resources, materials, etc.,and the search for appropriate methods and systems of accomplishing this.

In any case, it is not the novelty of the concept that gives it importance; it is, as Lieberman has pointed out, its utility as a unifying theme around which may be organized a number of the most prominent concerns on the current educational scene: systems analysis, operations research, performance contracting, even the voucher system and other freedom-of-choice plans. This umbrella aspect of the concept, by the way, accounts for the difficulty of providing any neat definition of the term accountability. It means many things to many people (a reason, perhaps, for its easy acceptance?). Yet, for establishment of accountability in any formal,systematic sense, certain common elements are discernible:

1. What are the schools to be accountable for? For student accomplishment and development - cognitive, affective, motor. This is taken to imply explicit, detailed statements of desired outcomes or goals, set forth in behavioral terms susceptible to observation or, preferably, measurement, in the absence of which statements there can be no evaluation of the enterprise.

2. Who shall be accountable? Our senses of logic and justice tell us that each person whose task it is to influence learning - teacher, supervisor, principal, curriculum coordinator, counsellor, whoever - should be held accountable for precisely that part of the educational outcomes which he can affect directly, through his own efforts. This highly specific imputation of responsibility is, as we shall see, a requirement which, if slavishly followed,nearly gives the whole game away.

3. How shall accountability be established? Clearly there is need for an accountability information system, providing systematic information on output and input. Further, there is need for a method for relating input factors, in- staff efforts, instructional materials, support systems, etc., to the

outcomes in a manner -- and here is the critically important point – that will permit the attribution of the outcomes in proper measure to the various input elements.

4. By whom shall accountability be determined? There is substantial feeling that, whatever a school or system may do in its own self-evaluative endeavors, independent auditors or "accounting" agencies are desirable.

By far the most comprehensive and sophisticated discussions of these elements of a formal accountability system that I have seen are contained in articles by Barro and Dyer in the December 1970 issue of the Phi Delta Kappan. Both papers manifest a very healthy awareness of the complexity of the data-gathering task and of the analytical methods that must be employed if it is to be possible to assign responsibility properly to the various contributing agents. Barro and Dyer have seen that it is extraordinarily difficult, perhaps impossible, to disentangle the several contributions of the variety of professionals to the learning of pupils. Dyer proposes that accountability always be thought of as joint accountability, by-passing, in a sense, any attempt to divide up the responsibility of various staff members and concentrating on responsibility at the school level through the creation of what he calls School Effectiveness Indexes. Both, interestingly enough, arrive at a multiple-regression approach as the appropriate analytical scheme, seeking thereby to partition the variance in student performance according to its various sources. Both approaches, it is interesting to observe, address themselves to the simpler case of a uniform set of outcomes for all learners. The alternative, and probably more nearly realistic case, in which desired outcomes are permitted to vary across learners, across classes, schools, and systems, is so very much more complicated as quite possibly to have seemed impossible to cope with at the present state of the art. In fact, if I read correctly between the lines of these two presentations, I sense a small

suspicion that even the models proposed may be seen by their authors as unattainable in the real world; yet, I take it that they are advocating that we must make the effort, whatever the likelihood of success - and if this is a proper interpretation of their sentiments, it accords with my own view of the matter.

Another attempt to establish accountability, somewhat in the Barro-Dyer model though less elegant and sophisticated, which has at least been pilot tested, is Project Yardstick, under the direction of Fred Pinkham in Cleveland. Yardstick provides a schema for relating test performance to several input measures. All three of these approaches, incidentally, are reminiscent of the New York State Quality Measurement Project of a decade ago, certainly the most sustained and ambitious effort of this kind. It would be instructive to know why it has not flourished.

THE PERFORMANCE CONTRACT

The Barro and Dyer models, in my opinion, point the way in which serious efforts at the assessment and location of accountability must proceed in the long run. In the meantime, in a more immediate ad hoc effort to establish accountability for certain aspects of the educational program, we have the phenomenon of the performance contract. In concept, the performance contract is simple: a school system specifies desired outcomes, defines the target group of pupils, establishes certain parameters within which education of the pupils is to take place, and enters into a contract with an agency for the provision of an educational experience that will bring the target group of pupils to the desired outcomes. The agency typically to date, though by no means necessarily, has been a private, commercial purveyor, leading to such witless treatment in the press as headlines which cry, "Can Big Business Succeed Where Schools Have Failed?" Let us hope we may be spared this red herring in considering the merits of performance contracting; the contracting agency may perfectly well be a university, a teachers'

organization, a professional society, or the like. Methods and materials are left to the discretion of the contracting agency, and it is hoped that they will bring innovative and extraordinary modes of instruction to bear. The contracting agency undertakes to bring about stipulated amounts of progress or improvement, generally but not always in a basic skill area, often agreeing to a penalty should pupils not reach the desired level, and, about equally often, stipulating that it shall have a bonus or premium for bringing the learners to a level in excess of the goal. The contractor, in other words, undertakes to insure gains or growth in accomplishment of a stipulated amount subject to penalties in the event of failure. The outcomes of the services provided under the contract are to be audited by an independent agency whose assessment of the amounts of gain that have taken place shall form the basis, in part at least, for the payment to be made to the contractor.

In the pure-culture performance contract – the Lessinger-Blaschke model — two additional features are regarded as central. These are the functioning of a so-called Management Support Group, which offers advisory assistance to the school with respect to the development of the specifications for the contract, the location of appropriate bidders, the award of the contract, and subsequent services; and the so-called "turnkey" provision, by which is meant the explicit inclusion in the contract of arrangements that will permit the methods, materials, practices, etc., of the contractor's intervention to be incorporated in the regular operation of the school or school system and carried forward by the regular personnel.

In much of the early discussion of performance contracts, there was reference to them as "guaranteed performance" arrangements, the implication being that the contractor warranted that every pupil would attain the stipulated goals. The guarantee notion seems to be less prominent lately, perhaps in grudging recognition

of a still-existing law of individual differences. The talk now is of performance contracts with premium and penalty clauses. (Speaking of guarantees, do you remember how last Sunday we all anguished with our astronauts as they struggled with the recalcitrant docking device? It is reported that Mission Control called the manufacturer of the device seeking his assistance and counsel, only to be told, "We're sorry; that unit has gone more than 50,000 miles and it's not under warranty any more." One wonders for how long some of the contract learnings may be guaranteed.)

A considerable number of school administrators, habituated to the purchase of a wide variety of services, such as maintenance, transportation, food services, etc.,on a performance-contract basis, responded enthusiastically to the notion that instructional services might be handled on a similar basis, particularly with respect to instructional problems that had resisted previous efforts. As Blaschke has repeatedly pointed out, the performance contract has appeal as a low-risk approach - low financial risk because of the guarantee or penalty clauses, and low political risk because failure could be imputed to the contracting agency rather than to the school system. The early favorable reports from the Texarkana project undoubtedly contributed to the eagerness of school men to explore and even to enter into such seemingly attractive arrangements.

But from the beginning there were also voices of caution, and not a few of outright opposition. Critics of the approach saw in it an abdication of professional responsibility, committing a part of the responsibility for the primary mission of the schools to nonprofessionals. Teacher organizations voiced particular concern on this score, in spite of assiduous efforts on the part of some of the contracting agencies to woo the support and collaboration of these groups. Some of the early advocacy of performance contracting was not without a note of hucksterism. Albert Shanker, President of the United Federation

of Teachers, denouncing performance contracting as a kind of educational "cure for cancer," declared that to guarantee performance in certain complex fields of human endeavor is to engage in deception. To "guarantee" to bring every child, regardless of ability, prior achievement, etc., up to the national norm seems to bespeak either a lack of awareness that the "norm" is by definition a level below which half of pupils in general achieve, or an extraordinarily, perhaps recklessly, high level of expectation. One should not begrudge a publisher or other contractor boundless confidence in his materials; but in the nature of things, not every pupil can wind up "above average."

Some publishers and other purveyors of instructional materials and support services regarded it as inappropriate for schools to seek guarantees of performance for the materials, since the purveyors had little control over the way in which materials and services were used. The notion that textbooks or other instructional materials could be "guaranteed" to produce specified amounts of learning struck many as reflecting a serious misapprehension of the nature of the learning process and of the role of the textbook -- as if a textbook had a definite, uniform, predictable impact on a pupil's learning, as a drug might on his body chemistry or physiological processes. Moreover, such research as is available on the contribution of the text or instructional materials to variance in pupil performance (as in the First-Grade Reading Study) suggests that this contribution is small - very much less than that of teacher competence, for example. Thus, many suppliers refrained from bidding on contracts where they could not exercise major control over the total instructional system, but were merely to provide materials. Those of us with long memories in textbook publishing remember when the harshest criticism of the textbook was its supposed straitjacketing or control of instruction and curriculum. Now it almost seems as if the instructional materials are to be required to display - guarantee - this monolithic impact on learning.

Other critics voiced uneasiness that the performance contracts would divert disproportionate amounts of resources to the pursuit of narrow and short-time goals, to the detriment of other objectives; some administrators felt that if the funds available for performance contracting could be channeled into their regular operations, they could accomplish as much or more as the performance contract arrangement. And some critics, lay and professional, took a dim view of the use of extrinsic motivators, such as trading stamps, radios, etc., employed, for example, in the Texarkana project.

The performance contract, as you all know, received its initial fame through the Texarkana project, and you are all aware of the melancholy fate that befell it at the end of its first year. The Office of Equal Opportunity has mounted a massive investigation of performance contracting, sponsoring performance-contract programs in some 18 school districts and arranging for a comprehensive evaluation of them. Meanwhile, it is reported that some 150 school districts have entered into one or another type of performance contract, covering a variety of programs over most of the elementary and secondary grades, with a wide range of conditions and through a sizable number of purveyors.

MEASUREMENT PROBLEMS IN PERFORMANCE CONTRACTING

The philosophical, political, and economic aspects of performance contracting are not my major concern tonight. They have been amply discussed in other forums, and I shall not dwell on them here, much less pass judgment on any of the issues in these areas. My mission is, rather, to invite your attention to certain of the measurement problems that inevitably arise in the conduct and, more particularly, the evaluation of a performance contract. These problems may be subsumed under familiar rubrics - validity, reliability, and unit and scalar properties of the measuring instruments.

Validity. The performance contract begins with a specification of the educational outcomes to be achieved through the contracted intervention; there is strong emphasis on the necessity for detailed enumeration of the behavioral objectives to be achieved. Under these circumstances, one would suppose that identification of appropriate instruments that would validly measure the attainment of these particular objectives would be greatly facilitated. So, indeed, it might - except for the overriding insistence that the results be expressible in units that are thought to be meaningful and comprehensible. This has eventuated, in the case of most performance contracts written to date, in a stipulation that the gains be measurable in terms comparable to "normal progress," generally defined as progress in terms of grade equivalents or, less often, age equivalents. This requirement has driven the contractors - reluctantly in some instances - to adopt one or another of the more widely used achievement series as the instrument for measuring gain, since these are the only series having dependably established normative systems yielding grade- or age-related measures. But these series are, almost in the nature of things, concerned with a much wider range of content and outcomes than the narrowly defined, more specific ones of the contract interventions, so that the fit between the goals of the intervention and the content or functions measured by the test leaves much to be desired. A considerable part of the variance in the scores

on these general achievement tests may be unrelated to the specific goals of the contract program.

There is a widespread belief among laymen (including legislators, in this context) and, for that matter, among a great many school people, that measurement of growth in reading ability can be satisfactorily accomplished through the repeated use of any of half a dozen of the better series of reading tests now available — and in a sense this is true. But everyone who is familiar with reading tests knows that the several reading tests do not correlate perfectly with one another, even within the limits of their reliabilities. The tests vary with respect to subtest composition, relative emphases on component skills, and so on; they may be equally defensible on rational grounds as samples of the reading domain, but it does not follow that each of them is equally valid or, indeed, that any one of them is valid as a measure of the particular reading objectives of a given performance contract. And as with reading, so with arithmetic and, to an even greater extent, so with the content areas of science and social studies. In a word, the nationally standardized tests on which performance contractors (or the evaluating or auditing agencies) have relied because of their credibility and their normative systems may, from a validity standpoint, be considerably less than ideal for the evaluative task.

Under the general heading of validity, I would like to dwell for a moment on the touchy issue of <u>teaching for the tests</u>. It is repeatedly suggested that the performance-contract type of arrangement, with its concentration on relatively narrow and specific goals, conduces toward instruction undesirably and narrowly focused on those behaviors that are the immediate target behaviors and which will presumably form the basis of assessment. Since most contract situations involve a pre- and a post-test situation, almost necessarily calling for use of alternate forms of a given instrument at the beginning and the end of the program, all concerned clearly are likely to know what the character of the final

assessment device will be, if not, indeed, to know precisely what its content will be. We need not be altogether cynical about human behavior to anticipate that this knowledge will condition and shape the pattern of some instruction. Such patterning may take the form of familiarization of the subjects with the actual exercises they will encounter in the final test, and we would say, ordinarily, with resulting contamination of the final test results and a subversion of any attempt at evaluation of gains. But the question is not quite so simple. At the early grade levels particularly, a performance-contract instructional sequence may be directed to the attainment of goals in realms where the universe of outcomes is limited. We may think, for example, of knowledge of letter names, a very early prerequisite for learning to read. The universe of outcomes consists of ability to recognize 26 lower-case letters and 26 capital letters. Here, clearly, the appropriat instruction program must consist of having the child perform precisely those behaviors that will be included on any test of his competence. The same is true of, let us say, mastery of the basic addition and subtraction facts, or of mastery of the spelling of the fifty most common words in primary reading materials. So we cannot say that any instructional practice in which a learner is exposed to precisely the tasks that he will encounter on a final assessment instrument is necessarily bad or wrong; but it is important to point out that the more specifically the desired goals are defined, and the more narrowly focused the instruction on these particular goals, and the more closely the post-test reflects and measures attainment of these goals, the more acute becomes the question of deciding what is and what is not legitimate approximation of test content and instructional content.

Reliability. On the matter of reliability, evaluation of performance contracts is particularly vulnerable to all the perennially vexing problems of the reliability of a gain score for an individual pupil. Even with tests having satisfactory reliability as measures of status of an individual pupil — say .90, out as high a level as is reached by most subtests in the commonly used batteries —

the reliability of gain measures over relatively brief periods, say four to six or seven months (the common duration of most of the early performance contracts), is distressingly low, influenced as it is by the measurement error in both pre- and post-test scores. The error of measurement of a gain score may very easily equal or exceed the amount of gain normally to be achieved in a short-term intervention. Yet is is on the basis of these gain scores that it is proposed that contractors be rewarded or penalized. It is ironic that whereas measurement textbooks caution against taking individual decisions or actions on the basis of measures having reliabilities of .4, .5, even .6, no one thus far seems to be very excited about making or withholding payments to a contractor on the basis of a piece of information of this degree of reliability. One recently announced contract, for a horrible example, "guarantees" individual gains of half a grade level in four months, for first-grade pupils, in social studies and science. To essay to discern, much less measure reliably, such differences in individual pupil attainment in these areas in grade 1, is really to wander in cloud-cuckoo land. (This same contract, by the way, calls for a bonus to the contractor for every child showing a significant increase in IQ. Well, why not?)

It is almost instinctive to react to this state of affairs by saying, "Well, let compensation be based on <u>average</u> gain for a group, and avoid the messy question of unreliability of individual gain scores." Such a proposal, acceptable though it might be to the contractor, is likely to be seen by the school as a cop-out -- and, I feel, not without some justice. It is clearly an intent of a performance contract to foster the academic growth of <u>every</u> participating learner, and no evaluation plan will be acceptable that allows failure by a significant fraction of the group to make good gains to be offset, in calculating payment, by better-than-average gains by others. The situation can be improved but not corrected by striving for greater reliability of both initial and final measures; for example, these measures might be based on administration of two forms

rather than a single form. But not only is this time-consuming and costly; it is not often the case that there are four equivalent forms of a measuring instrument available. Moreover, the increase in reliability of either initial or final status measures to be achieved by doubling the length of the test is modest, as is the reduction in the error of measurement of the gain scores from the lengthened measures. The more promising way of coping with this problem is to design projects of longer duration and not attempt to assess short-term changes, at least as a basis for compensation.

Alternate-form comparability. We have spoken of the use of alternate or equivalent forms as pre- and post-measures, and it is proper in this connection to observe that, even under the most conscientious test-building procedures, alternate forms may yield results, whether in terms of raw scores or converted scores, that are not precisely comparable. Determinations of equivalence of forms are necessarily specific to the equating sample and do not necessarily apply with equal precision to any other groups. Moreover, they involve necessarily their own sampling and estimation errors. The degree of imprecision is, in most uses of the tests, slight enough to be tolerable, but it can become important in a context where variations in compensation may turn on such minimally perceptible differences as a month or two of grade equivalent.

Level comparability. A similar situation prevails with respect to the equivalence of converted scores across successive levels of the more commonly used achievement series. For most of these series, the test development enterprise includes the articulation of successive levels, to permit translation of raw scores on the successive levels to some common set of units. Again, even when this is done with all conscientiousness, the precision of the translations can never be fully guaranteed. Contracts of the kinds entered into thus far may appropriately involve administration of different levels of a test at the beginning and end of the program; and the imprecision of the conversions may introduce additional distortions into vidual pupil gain scores.

Inter-test differences. Everyone knows, of course, that scores on standardized tests are likely to vary systematically from one test to another, as consequences of differences in their standardization groups, times at which they were standardized, varying content even in like-named tests, etc. Less well recognized is the fact that the various tests yield distributions of grade equivalents for given subjects at given grade levels that also differ systematically from one another. Test A, for example, will yield distributions of reading grade equivalents at grade 4 having larger standard deviations than Test B. The standard deviation of scores, whether raw or derived, is a function in part of the distribution of item difficulties and their intercorrelations, and is thus partly at the test-maker's discretion. Use of one or another of the available tests, accordingly, may produce different financial results for school and contractor, entirely as a consequence of this artifact and for no reason related to the effectiveness of the program provided.

The grade equivalent system. Faced with the financial consequences, to either school or contractor, of gains measured in grade equivalent terms, it is surely prudent to inquire whether the grade equivalent system is not too slender a reed to support such weighty baggage. One might have supposed that in 1970 practitioners of educational research would need no reminders of the limitations and deficiencies of grade equivalents, yet some of the practices built into performance contracts make one wonder. Grade equivalent scales are notoriously unequal-unit scales, having no zero points. They are most certainly not ratio scales. Thus, talk of "125% of normal gain," such as occurs in some performance contract language, is altogether meaningless; thus, gains of given numbers of months of grade equivalent represent accomplishments of quite different difficulty for a contractor to bring about according to the level and subject; thus, efforts to assess cost-effectiveness in any realistic sense are foredoomed, if output is measured in grade-equivalent terms.

The deficiencies of grade equivalents are particularly egregious in connection with the measurement of achievement at the secondary level. The Texarkana project, for example, sought, among other things, to raise the reading level of 9th-grade pupils by "one grade level." One has to wonder whether it was realized that a gain of one grade level, as measured by most reading tests for the secondary level, would correspond to a raw-score gain of not more than two or three points - in almost all cases well within the error of measurement of an individual score. The within-grade variance of scores on secondary achievement tests is so great in relation to the between-grades variance as to render grade equivalents altogether inappropriate. The logic underlying the development of the grade equivalent for secondary achievement tests is so irreconcilably at variance with the realities of curricular and instructional practices in secondary schools, and with the facts of student growth in academic achievement, that it is surely time for us to lay to rest this mode of interpreting - o should I say misinterpreting? - scores on secondary achievement tests.

Scarcely less unfortunate are efforts to interpret elementary achievement test performance exceeding the median performance of end-of-ninth-grade pupils by way of so-called "extrapolated grade equivalents" that purport to express performance of superior pupils in sixth, or seventh, or eighth grade as like the performance of typical 10th, or 11th, or 12th graders. Such extrapolated values are commonly identified by test publishers as artificial. Maybe it is now time to declare that their artificiality exceeds any potential utility and that they, too, should be quietly dispensed with.

And maybe the cumulative impact of all the problems enumerated above is sufficient to lead us finally to speak the unspeakable: to declare that the grade equivalent, at whatever level, is an inappropriate unit for the measurement of gain of an individual pupil over relatively brief periods — say as much as a year

of ordinary growth. Those of you who are familiar with the instruments that we publish, and their espousal of (though not exclusive reliance on) grade equivalent systems of interpretation, may be listening to me in wonderment and tempted to say, "Well, when did you kick the habit?" My answer to you, a little wistful perhaps, must be, "Not yet"; and please notice that my renunciation of grade equivalents is far from total. For all their limitations, they can, in my opinion, serve useful functions, particularly with respect to the assessment of progress of groups over longer periods of time. And if one asks, "Well, what better way is there, what better set of units for measuring academic gains?" we are hard put for an answer. We can point to efforts that have been made to develop continuous scales having units more nearly defensible as equal units, or to the utilization of within-grade status measures, such as percentile ranks or stanines, as bases for estimating magnitude of growth. For a variety of reasons we do not have time to go into here, these alternatives are considerably less than ideal.

The foregoing enumeration of technicalities will have seemed to many of you, I know, tedious, not to say boring; indeed, to those of you familiar with measurement, the cataloging must have seemed rudimentary. My justification for this discussion is that, in all the literature on performance contracting, I find few references to these measurement issues, and it has seemed to me worthwhile to get them into the record in a meeting such as this. The evaluation of performance contracts and the implementation of the accountability concept can ultimately be no more secure than the measurement data on which they rest; if these data are flawed by psychometric difficulties of the kinds I have suggested, we will never be in a position to assess properly the usefulness of this approach.

Criterion-referenced tests. I must not leave you with the impression that resort to grade-equivalent interpretation in performance contracts entered into thus far means that the contracting parties have been unaware of these difficulties. Confronted with the annoying metric characteristics of norm-referenced tests, they, and others, have sometimes sought to exorcise these demons by invoking the magic phrase criterion-referenced tests. There is much that might be said about the adequacy of this alternative, and much of that favorable. Certainly, strong arguments can be advanced to support the proposition that criterion-referenced tests might be more valid measures of certain performance-contracted outcomes. But it is not yet altogether clear how results of a series of criterion-referenced tests can be translated into units that will yield measures of gain or growth. This is not an impossible task conceptually; perhaps Rasch-model operations can point the way. Neither is it easily accomplished, nor can it escape many of the problems that we have enumerated above with respect to norm-referenced tests. Secondly, there seems to be an easy assumption that criterion-referenced tests of respectable quality and adequate scope can be called into being reasonably easily and quickly. The truth is quite otherwise. The methodology for development of criterion-referenced tests is less well explicated than that for the development of norm-referenced tests, but it is clear, to me at least, that the production of batteries of criterion-referenced tests equal in quality and scope to the better norm-referenced tests will be no mean accomplishment – and, in the long run, I suspect, not less costly than the development of norm-referenced tests covering essentially the same domain of knowledge and skills.

As an aside, the difficulties in arriving at any satisfactory estimate of "growth" of individual pupils, especially over short periods of time, have prompted some of us to have second thoughts about the entire concept of "growth" in academic attainment. The notion of "growth" in reading or arithmetic or language skills comes easily to us by ready analogy with growth in height or weight; but we may well

wonder whether the process by which a learner acquires more information in a given discipline or attains greater skill in, let us say, reading or arithmetic computation accords well with the model of growth in height or weight. We may further wonder whether tests built according to the methods used in constructing norm-referenced tests, having as their goal the maximization of individual differences, are efficient instruments for measuring this supposed growth – and we may ask ourselves how to define "normal" growth: normal for whom, under what conditions, etc. But these are speculations for another time.

Am I saying, then, that it is not possible to evaluate satisfactorily the outcomes of short-term interventions such as are called for in most performance contracts (having in mind, for the moment, only the psychometric considerations, and not other obviously relevant issues such as permanence of gains, Hawthorn effects, regression phenomena, comparison with control groups, transferability of the instructional programs and skills from the contractor group to the regular staff, etc.)? That is a rather harsher judgment than I am ready to make. I believe reasonably dependable estimates of average gains, at least in reading and arithmetic, can be obtained in most of the contract situations going forward at the present time, but I do not see a satisfactory answer to the question of sufficiently reliable measurement of individual pupil gains. Neither do I discern the logic that will permit a school system to ascribe even average gains unerringly to the contractor's performance or his special type of intervention. It would be my hope that performance contracts negotiated hereafter would contemplate intervention programs of greater duration than a few months, that greater attention be paid to reliability of initial and final measures, that the selection of evaluation instruments receive far more searching attention prior to writing the contract than I think has been true heretofore, and that far more comprehensive testing for formative evaluation purposes be built into the programs – systems of continuous performance monitoring.

What of the future of performance contracting? My crystal ball is as clouded as any man's, but that does not deter me from a little forecasting. Of course, we would all be well advised, would-be contractors and interested bystanders, to await the evaluation of the OEO-sponsored performance-contract programs now in progress before venturing to look too far ahead or to risk too much. My own opinion is that, within a couple of years, performance contracting as we now know it will be seen as a rather primitive, simplistic approach to the establishment of accountability. Blaschke has reminded us that the performance contract should not be viewed as an end in itself - that it is just one way in which a local school system may experiment effectively. Perhaps, given incentives such as those made available by performance contracts, local school systems may be motivated to seek change by other arrangements.

But whatever the fate of performance contracting, it is my feeling that the notion of accountability will continue large in our thinking, and a powerful influence on the educational scene. How can it be otherwise, when as a people we are committing such vast sums to education? I hope that our view of accountability will be a large one, that we will not permit ourselves to worry overmuch about paralleling precisely the accountability methods available to industry. For all our insistence on bringing every child "up to standard" in reading, arithmetic, etc., we still know that this is far from the whole of schooling. We know that education, unlike a manufacturing operation, must concern itself with raw material infinitely varied, and that it seeks a product, not of unvarying sameness as does the manufacturing operation, but with its initial richness and variety enhanced and multiplied. Who of us wants it otherwise? How to translate that richness and variety into behavioral objectives, how to assess their attainment in all their richness, and how to capture it all in cost-effectiveness equations, I do not know. But I believe strongly that even modest and limited successes are greatly to be preferred to faint-hearted failure.