Educational Evaluation and Policy Analysis
Fall, Vol. 8, No. 3, 1986, pp. 267-276

Assuring Quality in Educational Evaluation

William M. K. Trochim
Cornell University
and
Ronald J. Visco
The Providence, Rhode Island, Public Schools

Quality assurance methods are seldom used in educational evaluations despite ample evidence of serious research implementation problems and data errors. In this paper a number of quality assurance methods are illustrated using evaluation data obtained from the Providence, Rhode Island, School District. The methods are as follows: (a) from auditing, the method of internal control; (b) from accounting, the method of double bookkeeping; and (c) from industrial quality control, the methods of acceptance sampling and of cumulative percentage charts. Methods like these are useful for identifying poor quality research and targeting efforts to improve quality.

Educational data gathered at the Local Education Agency (LEA) level is usually plagued with high error rates, data losses, inaccurate program accounting, and other quality problems (Trochim, 1981). These are so serious that they often lead to questions about the interpretability of locally originated educational evaluations.

A serious barrier to quality is the lack of motivation to do good work (Boruch & Cordray, 1980; Boruch, Cordray, Pion, & Leviton, 1981; Boruch et al., 1982). Nevertheless, school districts that are motivated to improve research quality are without constructive suggestions from the methodological community about how they might proceed. A notable exception is the advocacy of Statistical Quality Control (SQC) techniques (Crane, 1979; Crane & Maye, 1980) in the Chapter I evaluation system. Even here, however, the emphasis has largely been on the detection of error rates at the State Education Agency (SEA) level based on sampling of district-level data.

In this paper we discuss several simple, relatively inexpensive quality assurance techniques that can be incorporated into LEA evaluations where the means and desire to improve research quality already exist. The techniques come from fields where issues of quality have long been a matter of contention: the auditing and accounting professions and industrial quality control. The application of these procedures to educational evaluation is illustrated, primarily using program and testing procedures from the Providence, Rhode Island, School District.

Auditing and Accounting

Most businesses use standard accounting and auditing practices to manage and guarantee the integrity of their financial information. Despite the long history of

these two professions for assuring quality, their procedures have rarely been applied to improve the quality of evaluation data (Hudson & McRoberts, 1984; Millington, 1983). In this section we describe one major application from each profession that might be useful in evaluation contexts.

Internal Control

Description. To assure quality of research it is necessary to have a clear understanding of what is to be implemented. This includes an overview of the entire research system and a description of lines of responsibility. In industrial and business contexts the delineation of the accounting system and the assessment of its quality are usually the role of the auditor. A major auditing task involves the determination of the "internal control" of the system in question (Arens & Loebbecke, 1980; Hermanson, Loeb, Saada, & Strawser, 1976). Typically, the auditor begins by dividing the whole system into subsystems or "transaction cycles" that are fairly distinct and more manageable for study. These might include the sales function, payroll and personnel, accounts receivable, accounts payable, and so on. Each transaction cycle is studied until its processes can be described well.

In most cases, a detailed flow chart showing how individual transactions proceed through the particular subsystem is the result of such study. The auditor must identify critical points on the flow chart and examine whether sufficient controls exist to ensure the integrity of the transaction information. An important guiding principle is the segregation of responsibility. For instance, those in charge of disbursing cash in an organization should not also have sole responsibility for the record keeping of those transactions.

Once the subsystem is understood, the auditor attempts to determine where control may be breaking down. This involves empirical investigations, often called *compliance tests*, to see where discrepancies may arise between the system ideal and its implementation. Depending on the circumstances, the auditor might sample records at several key points or follow single transactions through all or part of a cycle. The individual transaction cycles are linked and cross-cycle compliance checks are performed. The results of the internal control study are integrated with other information (e.g., the capabilities and quality of staff) into a final report that essentially states whether the auditor believes that the level of any error is within "acceptable" ranges.

Application. Procedures like this are almost never followed to ensure quality of evaluation although they are straightforward, economically viable (at least on a small scale), and obviously advantageous. Most school districts have no formal definitions of "data cycles" beyond a trivial or gross level of detail. Few have or could easily develop flow chart descriptions of their research system. Little attention is paid to who is responsible for what tasks, and seldom are check points established for assuring data integrity. Fewer still bother to monitor such matters except in informal, poorly recorded ways.

However, it is easy to envision procedures for conducting evaluation internal control studies. One must begin by describing manageable subcycles of the research process that can be studied separately. For instance, many educational evaluations, especially summative or outcome ones, can be divided well along methodological lines into six definable subsystems:

Sampling: The process of enumerating the population and correctly drawing the sample for the study.

Measurement: The process of administering any tests, interviews, or observational scales.

Design: The process of implementing the design. In most outcome research this primarily involves the process of assignment to program groups.

Program: The process of enacting the program or treatment condition (and comparisons, if appropriate).

Data Preparation: The process of preparing data for analysis, including any data exclusions, additions, aggregation, and index formation.

Analysis: The process of analyzing the data, either quantitatively or qualitatively.

These subsystems offer a general framework for analysis but should not be applied literally in all cases. For instance, in many educational evaluations, sampling

from a population is less relevant than accurate measurement. Further, one need not be as concerned about defining the category into which some research step falls as in assuring that all important steps are included somewhere. The process of scoring multiple-choice tests, for example, might be equally appropriate under either measurement or data preparation.

Illustration. To illustrate the flow-charting task in this context, we can examine the system that is used to conduct Chapter I compensatory education evaluation in the Providence, Rhode Island, School District. Figure 1 shows a gross overview of all the subsystems involved. The feedback loop between the program and measurement subsystems indicates that repeated tests (pre and post) are administered.

Sampling issues are minimal (data are generally recorded for entire student populations), but serious implementation questions arise in all of the remaining subsystems. As an example, focus attention on an overview of the measurement subsystem as shown in Figure 2. The subsystem is divided into three major areas.

In Figures 3a–3c, the three measurement areas are broken down even further. Here, one is at a level of specificity that is probably sufficient for internal control in education. Each box on the chart refers to some checkpoint question that could be examined to determine system integrity. A first step in determining control in this chart would be to identify the person or persons responsible for each question. For example, the question "Were correct materials received?" involves checking by a clerk in the central evaluation office; "Were materials correctly distributed to schools?" is the joint responsibility of the schools and the central evaluation office; and so on.

The principle of segregation of responsibility implies the types of controls that might be useful at various points. For instance, if one person is responsible for one or more boxes in a row, it would probably be useful to have another individual, preferably from another department, to sign off on the work. It might be advisable to have a designated official in a school finance office (rather than the designated in-school evaluation supervisor) sign off on whether the correct tests were received or whether they were packaged correctly for scoring.

When multiple responsible agents are assigned to a single task, one can look for possibilities for redundant record keeping (or double bookkeeping as described below). In many educational settings, a good

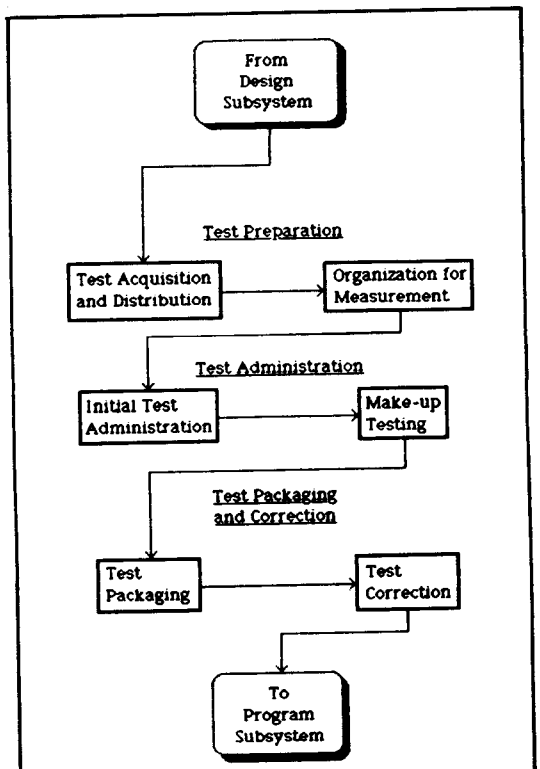

Figure 2. *A sample measurement subsystem in an internal control study in an educational evaluation context.*

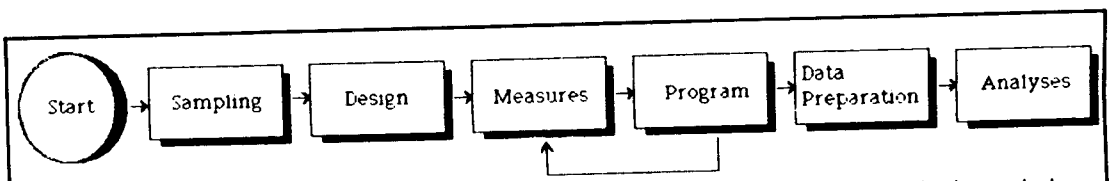

Figure 1. *Overview of possible subsystems in an internal control study in an educational evaluation context.*

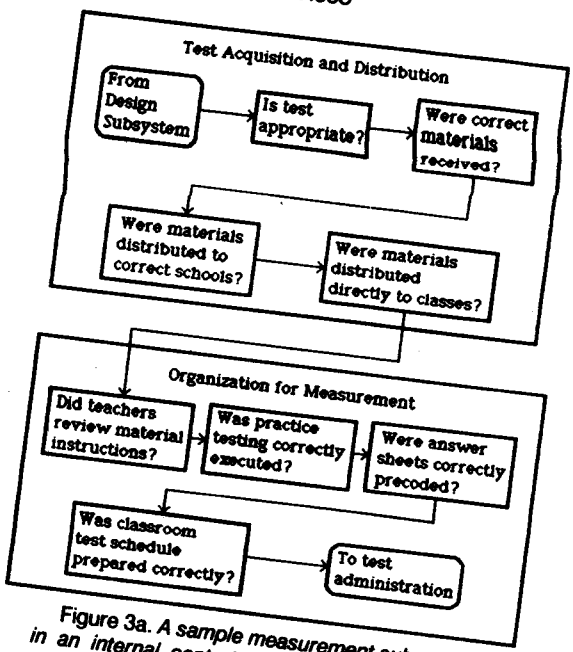

Figure 3a. *A sample measurement subsystem in an internal control study in an educational evaluation study: detailed view, Part 1.*

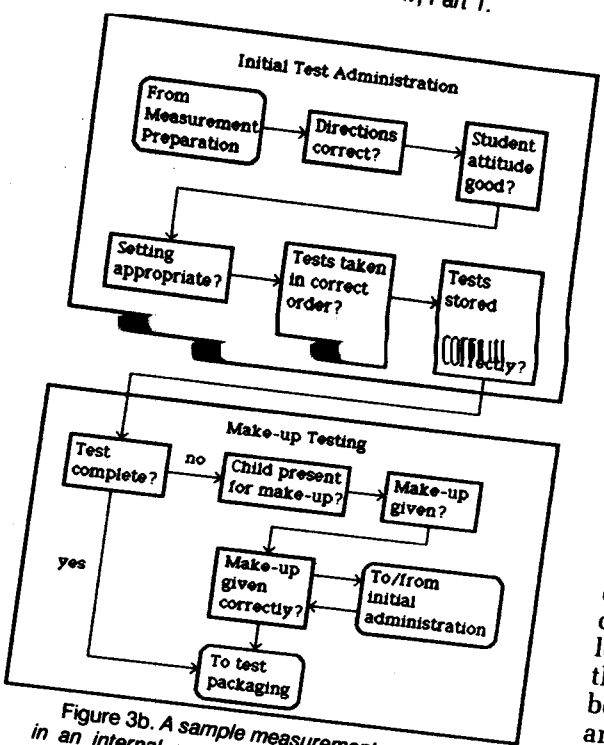

Figure 3b. *A sample measurement subsystem in an internal control study in an educational evaluation study: detailed view, Part 2.*

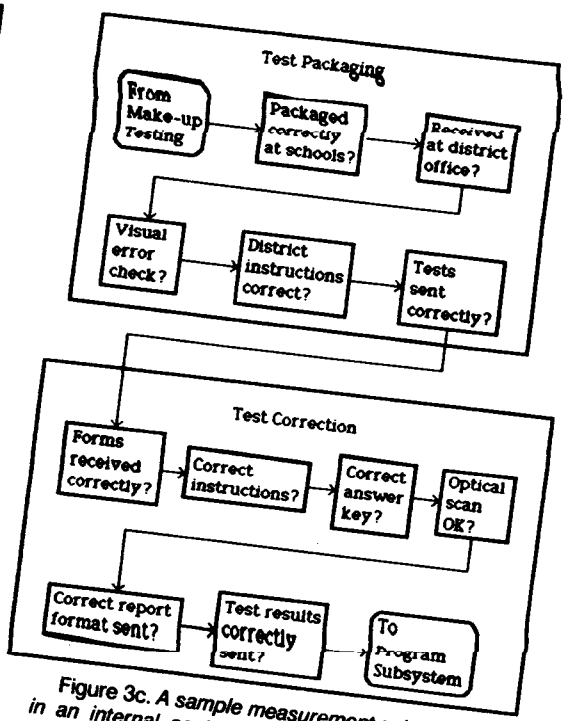

Figure 3c. *A sample measurement subsystem in an internal control study in an educational evaluation study: detailed view, Part 3.*

first step toward quality would involve the clear delineation of responsibility and the construction of "sign-off" points at key places of the flow chart. At the least, this would force the evaluator to be clearer about the specific criteria for acceptability and would increase the awareness and public accountability of the responsible agents.

Once the system is understood, one can begin to check its compliance with these criteria for quality. At some points in the measurement subsystem, for example, it would be efficacious to select a sample of tests or classrooms for spot inspections of quality. At other points, one might be able to review the entire population. For instance, a key source of errors involves the coding of classroom and school test package cover sheets prior to scoring. A single error here could cause all subsequent records to be sorted into the wrong grade level or program condition. Here, given the relatively few header sheets, it would be reasonable to conduct a thorough examination of all of them.

Double Bookkeeping

Description. The accounting profession is devoted to assuring the accuracy of financial information and uses strategies that could help improve research quality. Tabor (1977) provides an excellent discussion on evaluation data corruption from an accounting perspective. A major accounting procedure—double book-

keeping—has considerable potential for improving the quality of research in education.

Double bookkeeping strategies are probably the single most important methodological tool in the accounting profession. The use of double bookkeeping approaches can be traced back at least to 13th- and 14-century Italy and may in fact have originated in ancient Rome and Greece (Peragallo, 1938). The essential idea involves keeping two independent records of each financial transaction. Typically these ledgers are labelled credit and debit or income and expenses. The built-in redundancy of a double bookkeeping system guarantees that there will be some independent information gathered on the accuracy of the data. Accounting techniques have been used in education for financial purposes but have seldom been applied in evaluation settings.

Application. Consider the application of double bookkeeping procedures to one of the most difficult problems in educational evaluation: determining who received the program or treatment and for how long. Typically, students are added to or dropped from programs throughout the school year, even when the dictates of good evaluation would indicate that it is not advantageous to do so. In addition, most school districts have standard procedures for "challenging" students into or out of the Chapter I program on the basis of teacher, parent, or administrator views that the child was "incorrectly" assigned. The tendency to lose track of which students are receiving the program is great, particularly if record keeping is poor. Especially in areas where mobility is very high, record keeping that is typical in education has a considerable error rate. Usually districts take a census at some regular interval that is a "snapshot" view of school membership. On any given day in a school year it would be difficult to say how accurate the census remains. A double bookkeeping system could be used to improve ongoing record keeping between censuses.

Illustration. The potential utility of double bookkeeping can be illustrated with the approach used by the Providence school district for keeping track of which students are receiving Chapter I instruction in each school. Suppose that a student in a Chapter I program moves from School *X* to School *Y*. School *X* completes a "drop" form and submits it to the central Chapter I office, indicating that the student is no longer being serviced. As soon as the student arrives, School *Y* sends an "add" form to the central office, indicating that the student is now being serviced there. Occasionally one of the schools is remiss in sending in the appropriate form; a school might not send in a form until considerable time has elapsed, if at all. However, the double bookkeeping approach requires that both schools send in forms. If only an add form is received for the student from School *Y*, the central office immediately contacts School *X* to verify that the student was dropped, and requires the form to be sent. On the other hand, if an add form has not been received within about 2 weeks after receiving School *X*'s drop form, School *Y* is contacted to verify that the student had been added, and to request that the form be sent.

Suppose, however, that neither school sends the appropriate form to the central office. To guard against the effects of this double failure, an additional system check has been established. Each school's monthly attendance report is compared to the previous month's report. This procedure reveals any discrepancy in the list of program participants based on central office records compared with school records, and particularly whether students have been added or dropped without forms being submitted. In this event, both schools are contacted for verification.

Quality Control

Quality control and the correct implementation of production processes have long been central concerns in industry. Most moderate-sized industrial firms have a definable unit whose sole responsibility is quality control. Over the past few decades techniques have been developed to assess and improve quality in this context.

Acceptance Sampling

Description. One quality control strategy that would seem to have direct applicability to educational evaluation is termed *acceptance sampling*. Although technical discussions of acceptance sam-

pling techniques can become quite complex (Grant & Leavenworth, 1980), the essential principle is simple to understand. Acceptance sampling begins with the concept of a "lot" or group of data. In a school, for instance, the data from classrooms might be considered lots. At the district level, the data reported from each school can be considered a lot. In acceptance sampling, one does not examine every part of every record for quality. Instead, a sample of each lot is randomly selected and the entire lot is accepted or rejected, depending on whether the sample meets predetermined criteria. The decision of whether to use acceptance or 100% sampling is essentially a trade-off between cost and accuracy. With 100% sampling one tends to get higher accuracy (although fatigue factors may come into play), but such extensive inspection increases the costs.

Application. An excellent example of the use of acceptance sampling techniques can be found in Crane and Maye (1980), who use sampling strategies to assess the frequency of "correctable" data errors in Chapter I compensatory education evaluation at the SEA level. Using a sample of the results reported to the state of Illinois for several years, they demonstrated that previously undetected errors resulted in a positive bias in the state-aggregated estimate of program effect. Furthermore, their analysis indicated which of several types of errors were most prevalent and deserved the greatest attention in any data correction scheme. Finally, on the basis of their sample they estimated the average time required to correct each error type and used these time estimates in planning a cost-efficient data checking strategy. The sampling theory used in this study may be beyond the level of sophistication of many school districts. Nevertheless, simple acceptance sampling procedures could be adapted for use in a more local context.

Illustration. An acceptance sampling plan is applicable at several points in the Providence school district evaluation process. For instance, it would seem especially valuable when the district evaluation office does its check of test answer sheets before sending them to the correction service. This is a time-consuming process and, in part because it takes so long to inspect the answer documents, there is typically no time left to return forms with errors to the school level for correction or verification. As a result the evaluation office tries to make corrections in addition to doing the inspection.

The advantage of an acceptance sampling procedure to the district office is evident. In very little time, answer sheets from each school, or even from each classroom from within the school, could be sampled and the results compared to some set criteria. For instance, one criterion might be to allow for no more than 5% of the sample of answer sheets to have errors in coding student ID numbers. Multidimensional criteria would be even more desirable, although they would take longer to apply. The lots that failed the test would be returned to the school or classroom level for correction.

This direct and immediate feedback capability to the school and classroom is sorely lacking in most education evaluations. Typically, because of time constraints and the need to get the tests corrected quickly, classroom teachers don't find out about gross errors until after evaluation results are already compiled. By then, it is obviously too late to do anything about them. One way to encourage good work would be to make certain that the procedures allow time for immediate feedback on performance.

Cumulative Percentage Charts

Description. The cumulative percentage chart idea stems from what is often termed the Pareto principle, which is described in Juran and Dryna (1970):

> Despite the drama of sporadic troubles, the great majority of conformance losses are found to exist in a relatively few chronic troubles. This phenomenon arises from the invariable 'maldistribution' of defects. (pp. 9–10)

To construct a single cumulative percentage chart one needs data on the frequency of some problem or error across a number of units or samples (e.g., schools, classes, grades). The frequencies are sorted in descending order so that the unit with the greatest amount of error is listed first. The sorted frequencies are converted to percentages of total error, and the cumulative

percentages are then computed by summing down the percentage column. When these cumulative percentages are plotted, the graph describes the cumulative errors (from most to least) across units.

Although a single chart of this type may be useful, the cumulative percentage chart idea has greater value when multiple problems are graphed together. To accomplish this, one begins by selecting a "primary" problem area (usually the most serious source of error) and constructing the cumulative percentage chart as described above. For each other problem area of interest, one first constructs the percentage of total error for each unit. Then these percentages are sorted according to the unit order obtained for the primary problem measure. Once sorted, the cumulative percentages are then calculated for each of these additional problem measures. Usually it is desirable to plot all of these cumulative percentage charts on the same graph so that the relative error contributions of the units on a number of different problems can be assessed simultaneously.

Application. These charts taken together allow the evaluator to plan for quality control programs multidimensionally. In a situation where there is enough money for implementation of a general quality control push in only five schools, the evaluator can select the five highest on the primary problem measure; it will be known in advance not only what percentage of that problem is likely to be addressed but also what percentage of other relevant quality control problems will be attacked by a general quality program at those five sites. This easily accomplished technique is simple to apply, readily understood, and goes a long way to help target areas that need to improve implementation.

Illustration. One of the most serious problems in the Providence evaluation cycle occurs when the test answer documents are coded with demographic information and item responses at the school. Each student is asked to "code" his or her name, school department identification (ID) number, test level and form, date of birth, and so forth, on the front of the answer document; teachers or aides sometimes assist with this task. Of course, during the actual testing, students code their responses to the items.

Unfortunately, the inaccuracy of coding is a major concern. The extent of the problem becomes clear when the answer documents are received at the evaluation office to be reviewed and packaged for submission to the test scoring service. The evaluation office examines every answer document for any apparent coding errors. For example, when coding ID number, a student may darken more than one of the "bubbles" within a single column, begin coding in the wrong column, or skip a column by mistake. Since the ID is so important to subsequent evaluation procedures, such as student matching, the office looks for these and other noticeable errors.

During a recent test administration, a tabulation was made of the type and frequency of coding errors that occurred within each grade at each school (testing is done in grades 1 through 6). Figure 4 depicts the actual distribution of ID coding errors, as represented by a cumulative percentage chart. The dashed line depicts the cumulative percentage across all grades; the solid lines provide figures for individual grades. It may be seen that School 1 (out of 22 schools) accounted for over 25% of all ID coding errors detected. The first five schools together were responsible for over 60% of ID coding problems.

A quality control program to address ID coding errors should do well even if it is limited in scope to these five schools. Of course, an alternate approach would be to address ID coding problems at specific levels within certain schools. For instance, Figure 4 shows that about 65% of all grade 6 ID errors occurred at School 1. A quality control program at School 1 would be more time and cost effective if it were concerned only with grades 6 and 3, where the great majority of ID errors occurred; a program implemented at the other grades would have less to accomplish. By addressing only critical grade levels within identified schools, a quality control program could be implemented at a greater number of sites.

Figure 5 depicts cumulative percentage figures for all coding errors (including ID) from the same test administration. Note

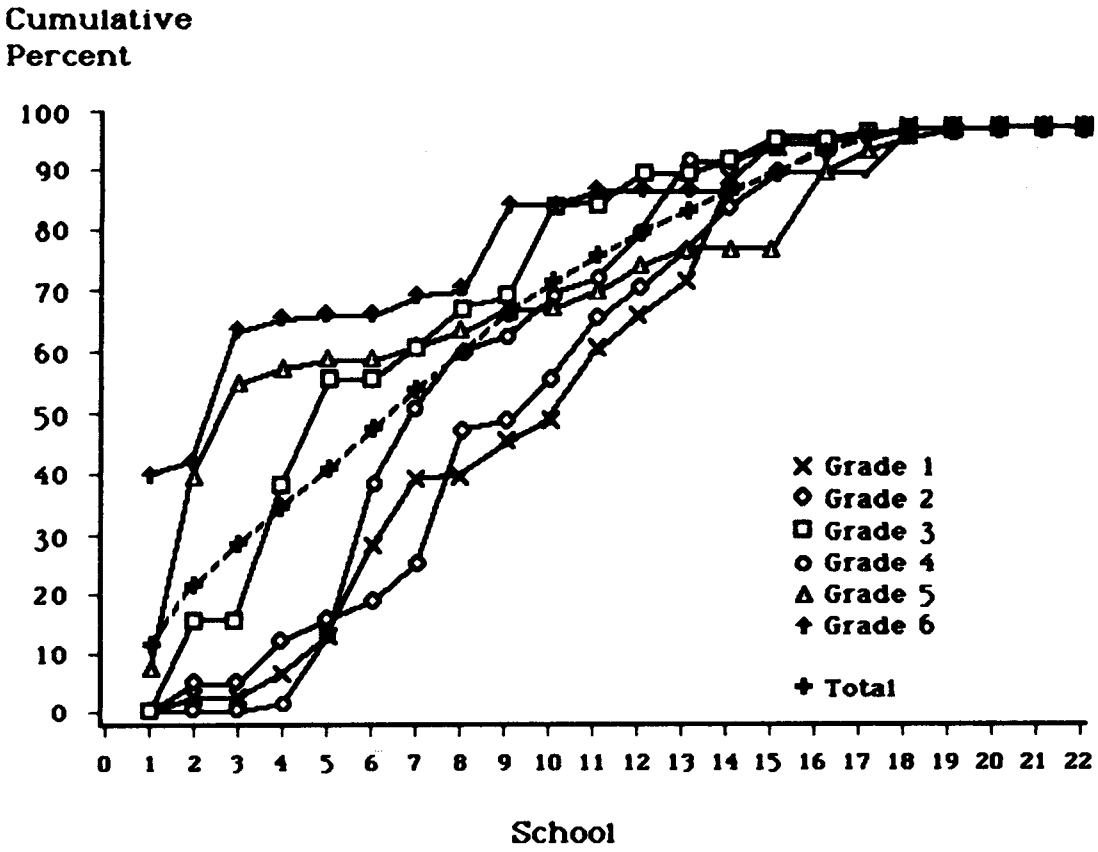

Figure 5. *Cumulative percentage of total errors by school for grades 1 to 6 and total, Providence Rhode Island School District.*

one might look at whether this reduces overall error rates more than checks conducted at the central district office. If the costs of each procedure can be estimated, one could compare the relative cost efficiency of the two methods for reducing error.

Along with testing the relative advantages of alternative control procedures, there is a need to gain greater experience in applying the strategies in educational contexts. We are particularly naive to the social and political implications of trying to improve evaluation quality, although we are sensitive enough to the possibility of difficulty to approach the issue cautiously. To the extent that our strategies require additional workloads, compete with program funds, threaten individuals, or make programs look worse, we can expect resistance from the system. Empirical examinations of attempts to institute quality control programs would help clarify major points of resistance and perhaps suggest alternative approaches.

In this paper we deliberately focused on only a few strategies that might be used to improve the quality of educational evaluation. The illustrations presented here have shown that some of these techniques have already been implemented in a major school district and that still others are feasible in that same context. Other strategies exist, some of which are more familiar to the educational community, but few of which are currently used to achieve quality assurance (Trochim, 1982). Before we can move ahead with concerted attempts to improve quality in educational evaluation we need to review the range of options available to us and place them in the context of the motivational and resource issues that could reduce or nullify their impact.

References

Arens, A. A., & Loebbecke, J. K. (1980). *Auditing: An integrated approach* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Boruch, R. F., & Cordray, D. S. (1980). *An appraisal of educational program evaluations: Federal,*

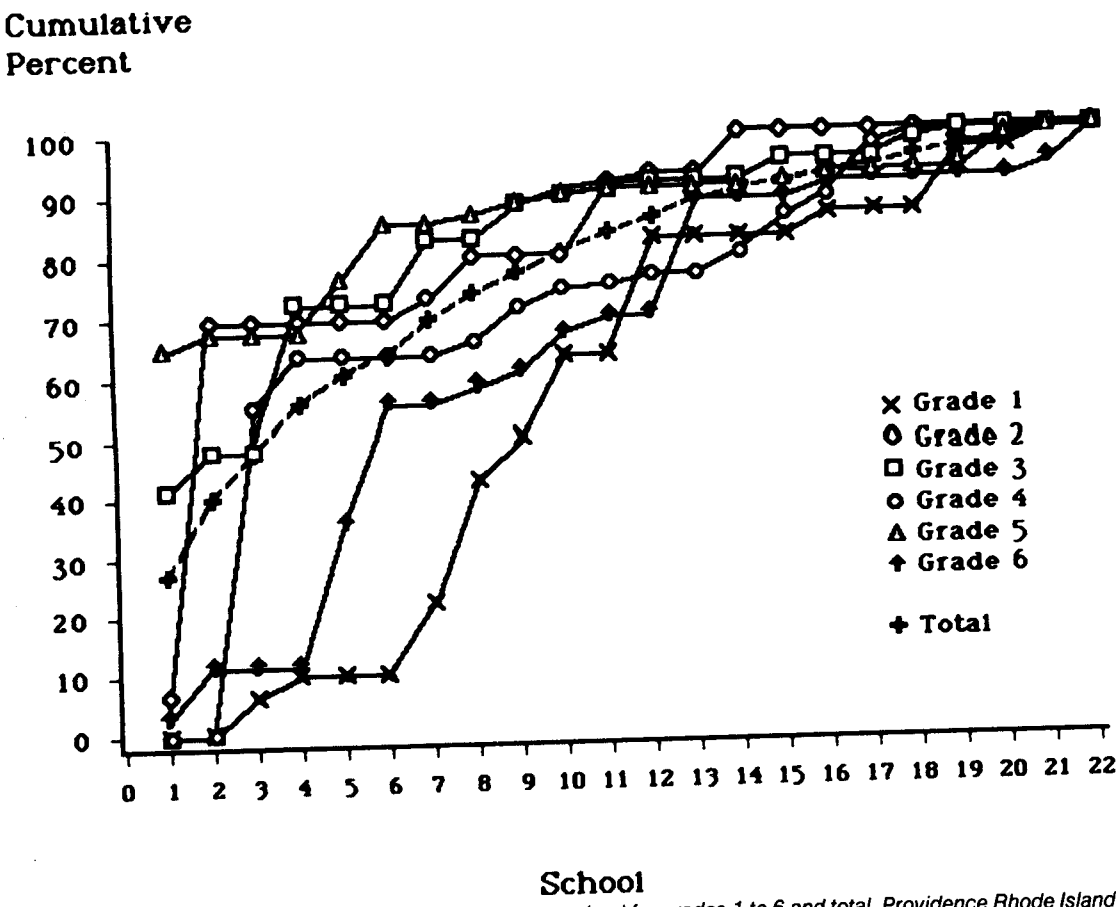

Figure 4. *Cumulative percentage of ID coding errors by school for grades 1 to 6 and total, Providence Rhode Island School District.*

that the school numbers in Figures 4 and 5 do not necessarily correspond, since schools are being ordered on different error variables. School 1 in Figure 5 accounts for about 12% of all coding errors, but grade 6 in School 1 accounts for almost 40% of all sixth grade coding errors. This result reinforces the desirability of aiming a quality control program at specific grades within identified schools. Since the evaluation office places a priority on the accuracy of student ID coding, that area would be the focus of most quality control procedures. However, at sites (and grade levels) where all types of answer document coding are a clear problem (such as Grade 6 in School 1 from Figure 5), a broader quality control focus must be attempted.

Discussion and Conclusion

The notion of quality in educational evaluation or any other endeavor is conceptually related to the value of the product on which quality is desired. Because quality is typically a relative ideal, one must be concerned about the amount of benefit that would be accrued with an incremental increase or decrease in quality. The value of evaluations of different quality, whether stated in dollars or some other currency, is difficult to determine and is likely to remain so. Therefore, questions about gains in quality that might result from the application of quality assurance techniques might best be phrased in relative terms.

Rather than asking whether an increase in evaluation quality as a result of some procedure balances the costs of the control mechanism, we might prefer to ask whether the technique can reduce costs of conducting evaluation somewhere else in the system. For instance, if a district decides to institute quality criteria checklists that classroom teachers complete to vouch for the quality of test information,

state and local agencies. Washington, DC: U.S. Department of Education.

BORUCH, R. F., CORDRAY, D. S., PION, G., & LEVITON, L. (1981). A mandated appraisal of evaluation practices: Digest of recommendations to the Congress and to the Department of Education. *Educational Researcher, 31,* 10–13.

BORUCH, R. F., CORDRAY, D. S., PION, G. M., SPENCER, B. D., TROCHIM, W. M. K., & WICK, J. (1982). Changes in evaluation practice at local and state levels and the possible influences of the Chapter I evaluation reporting system. In E. R. Reisner, M. C. Alkin, R. F. Boruch, R. L. Linn, & J. Millman (Eds.), *Assessment of the Chapter I evaluation and reporting system* (pp 67–84). Washington, DC: U.S. Department of Education.

CRANE, L. R. (1979). *Statistical quality control applications to Chapter I evaluation data.* Evanston, IL: Educational Testing Service.

CRANE, L. R., & MAYE, R. O. (1980, April). *Effects of correctable errors on Chapter I NCE gain estimates and implications for statistical quality control.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

GRANT, E. L., & LEAVENWORTH, R. S. (1980). *Statistical quality control.* New York: McGraw Hill.

HERMANSON, R. H., LOEB, S. E., SAADA, J. M., & STRAWSER, R. H. (1976). *Auditing theory and practice.* Homewood, IL: Richard D. Irwin, Inc.

HUDSON, J., & MCROBERTS, H. A. (1984). Auditing evaluation activities. In L. Rutman (Ed.), *Evaluation research methods: A basic guide* (2nd ed., pp. 219–236). Beverly Hills, CA: Sage Publications.

JURAN, J. M., & DRYNA, F. M., JR. (1970). *Quality planning and analysis.* New York: McGraw Hill.

MILLINGTON, P. S. (1983). *Internal audit as evaluation.* Unpublished manuscript, Stanford University.

PERAGALLO, E. (1938). *Origin and evolution of double entry bookkeeping.* New York: American Institute Publishing Company.

TABOR, J. G. (1977). The role of the accountant in preventing and detecting information abuses in social program evaluation. In H. W. Melton & D. J. H. Watson (Eds.), *Interdisciplinary dimensions of accounting for social goals and social organizations.* Columbus, OH: Grid, Inc.

TROCHIM, W. M. (1981, October). *Research implementation.* Paper presented at the Annual Conference of the Evaluation Research Society, Austin, TX.

TROCHIM, W. M. (1982). *Research implementation: A final report to the National Institute of Education.* Washington, DC: National Institute of Education.

Authors

WILLIAM M. K. TROCHIM, Associate Professor, Department of Human Service Studies, Cornell University, Ithaca, NY 14853. *Specializations:* Research methodology, program evaluation.

RONALD J. VISCO, Computer Education Administrator/Evaluator, Providence School District, 480 Charles St., Providence, RI 02904. *Specializations:* Measurement, evaluation, computer applications in education.