Implementing Performance Monitoring: A Research and Development Approach

Gary T. Henry, Georgia State University
Kent C. Dickey, Virginia Department of Education

How can a useful performance monitoring system be designed? Gary T. Henry and Kent C. Dickey describe the use of a research and development model to develop an effective performance monitoring system as part of the State of Virginia's efforts in educational reform. They provide details on how the approach was implemented and discuss both the limits and risks involved.

A combination of public interest, political pressure, and increased technical capacity has thrust public accountability, and with it the concept of performance monitoring, once again into the fore. When the U.S. Department of Education sounded an alarm with *A Nation at Risk* in 1983, the nation renewed its concern for public education. The presidential Educational Summit in Charlottesville, Virginia, the advent of significant community and parental control of Chicago's public schools, and the court-ordered complete overhaul of the public school system in Kentucky amplified the alarm. Dissatisfaction with public education services in many quarters has led to a variety of proposals ranging from dismantling educational bureaucracies by allowing school choice to increasing the accountability of the schools to the public.

A common thread woven through most of the proposals is the need for a performance monitoring system, also referred to as an educational indicator system. Performance monitoring is unique among the proposed educational reforms in terms of its duality of purpose. The first role for educational performance monitoring is as a part of the overall reform package. Performance monitoring provides periodic information on the outcomes and impacts of public services (Wholey, 1979; Poister, 1983). Teachers and administrators can use information on educational performance as an empirical base for observing strengths and weaknesses of educational practices and programs. The empirical baseline can then be used by stakeholders for planning and setting specific objectives for improvement, an important factor in realizing educational improvements (Ford Foundation, 1972). Other uses of the education performance data, such as incentive-based rewards, are also included in many reform recommendations (Richards and Shujaa, 1990). The first role for using performance monitoring data is using it as a part of the reform package; that is, as means of facilitating educational improvement.

In its second role, performance monitoring provides a means for holding a mirror up to the reforms. It serves to measure the effectiveness of the reforms that have been undertaken. Wohlstetter (1991) describes performance monitoring as asking the simple question, "Are state education reforms improving schools?" Performance monitoring can show the results and context of schooling to educators, parents, politicians, business representatives, and the public. The indicators of the "collective well-being" of education can communicate essential information and "add depth and breadth to the public's understanding" (Morgan, 1991, pp. 10-12).

The difference between these purposes is one of planning versus evaluation. In the planning function, the system is used as a means for improving education. To serve the evaluation function, it is called upon to assess the impact of the improvements. Together the functions can provide greater accountability and a basis for actions to improve education.

Accountability, Performance Monitoring, and Bureaucracy

Accountability for the results of public programs has become a concentrated focus of concern from elected officials and the media. In many policy areas, interest in accountability is growing (Benveniste, 1985). In the field of education, a sense of frustration is growing among politicians, policy makers, special interests, and the public who express the belief that ample time has passed for education reforms of the 1980s to have had an effect. They want to see results. Educational performance monitoring systems are the means proposed to show them the results.

In 1987, 21 states reported having an educational performance accountability system that linked performance results to policy actions (Office of Educational Research and Improvement, 1988a). In addition, more than 30 states had educational accountability on the agenda at the beginning of the 1989 state legislative sessions (Pipho in Wohlstetter, 1991). The movement toward greater public accountability and the specific dissatisfaction with education have wrested control of the issue of whether educational programs should be evaluated from "the satrapy of educators" (Timar, 1989, p. 2).

Instituting a performance monitoring system presents a practical problem for the educational administrators who receive the charge: How can a visible innovation such as a performance monitoring system be implemented in a public organization? We argue that the approach should be based on a research and development (R&D) model. An R&D approach to performance monitoring, while creating some difficulties due to the openness and visibility of public sector innovations, acknowledges the current state-of-the-art in performance monitoring. Not enough is known about measuring program outcomes and establishing objectives that will motivate higher performance levels to install a complete, operating system. Support for this assertion is indicated by the difficulties experienced with existing performance monitoring systems. Early leaders in the development of educational performance monitoring systems, California, New Jersey, and Delaware, are undertaking significant changes in their education performance monitoring systems (Goldman, 1990).

Performance Monitoring Systems: An R&D Approach

The research and development approach in the public sector is defined by four characteristics. First, the R&D approach requires that multiple studies be undertaken as the program is being planned to assess specific components of the program. Second, judgments about the way in which components are to be implemented are withheld until data are available. Third, program personnel must use empirical data to make decisions on program components and the alternative ways the components could be implemented. Finally, the testing of the program components goes on in increasingly "real" circumstances, in essence moving the project from the laboratory to the field.

The research and development approach does not question whether or not the program will be implemented. Rather, it seeks to bring empirical information to bear on the viability and utility of specific program components. By focusing on the components, research questions can be broken into researchable chunks and data addressing the questions can be sought. The central question about the program involves the "go—no go" decision. This decision has usually been made in the political process. The task posed to administrators is how, not if, the program is to be set up. Often, little direct experience, theory, or empirical data exist to guide the implementation. Expert opinion, intuition, and perceptions of political realities will define how the program will be implemented. The R&D approach offers improved probability of program success and reduced probability for problems.

Establishing performance monitoring systems through a research and development process may eliminate some longstanding flaws identified by Wilson (1989): incorrect and irrelevant data, untimely and inaccurate reporting, and lack of management support. Another value of the research and development model is that real or perceived negative impacts can be prevented by publicly emphasizing the opportunity for trial-and-error adjustments based on empirical evidence. It is an alternative to the innovation model behind the large-scale social experiments of the 1970s and wholesale changes such as the welfare reforms and workfare policies instituted in the 1980s.

Establishing a performance monitoring system can best be undertaken as a research and development enterprise going on in a public, political arena. Research, development, and innovation take place over a range of activities, including institutional arrangements, selection and measurement of indicators, and reporting formats—all within a diverse, volatile political arena. The developmental, high-stakes nature of the enterprise, the changing political environment, and the uncertainty of the results of various aspects of the system explain the necessity for modification and change such as those in the three states mentioned above. The R&D approach allows administrators to acknowledge this upfront and view the process of adapting performance monitoring as knowledge and understanding increase. The characteristics of public programs that are appropriate for using an R&D approach are shown in Table 1.

Table 1
Characteristics of Programs Benefiting from R&D Approach

General Characteristics	Performance Monitoring Context
Lack of experience with or knowledge of program components	Performance monitoring is a relatively new approach in developing, collecting, reporting and using educational outcome data.
High probability of unanticipated consequences	Indicators can provide incentives for questionable program practices and resource allocations.
High costs of program failure	Considerable political capital and credibility of education administrators at stake if program languishes or fails.
High degree of anxiety among program personnel	High stakes innovation due to uncertainty of results. and potential for misuse of educational outcome data.

Wholey, one of the original proponents of performance monitoring, conceived of a system that "measures inputs, activities, and outcomes and compares program performance with prior or expected performance" (1979, p. 117). Poister (1983), concentrating on state and local systems, developed requirements for a monitoring system that would utilize some of the traditional input and process measures and strengthen them by emphasizing "evaluative measures, especially those related to the program's outcomes" (p. 3). Both authors identified similar models for developing performance monitoring systems as shown in Table 2.

These two models of performance monitoring systems can be distilled into four distinct, but interrelated components, which can be used to guide research and development activities :

- Establishing objectives;
- Selecting performance indicators;
- Establishing performance criteria and methods for comparing performance; and
- Reporting and using performance information.

Each of these components is discussed below in terms of the R&D approach taken for an educational performance monitoring system in Virginia. In this article, five research activities are highlighted to demonstrate the use of an R&D model. An overview of the approach is presented in Table 3 and the five research activities are summarized in the next sections of the article.

Stakeholder Group Participation

A unique approach was used in the development of the components of the system, such as establishing objectives and selection of measures: the involvement of stakeholder groups in designing the system. The literature on performance monitoring emphasizes the need to involve managers and policy makers in decisions about the design (Wholey, 1979, pp. 61-77; Poister, 1983, p. 10). The educational environment offered the opportunity to extend beyond the managers to seek input and guidance from others involved in education. For this example, four groups were selected to provide advice and

Table 2
Performance Monitoring Logic

Wholey's Model **Resource Allocation to Evaluation**	**Poister's Model**
Select Program objectives Measures of performance Intended uses of performance information	**Management** Program objectives Program elements Performance indicators Intended uses
Performance Monitoring Activities	**Data Component** Data collection Data processing
Performance Monitoring Products Measurement of program performance Comparison of actual program performance with expected or prior performance	**Analysis Component** Measurement of current performance Comparison of current performance with Past levels Among units Against plan
Use of Performance Information Maintain or change objectives Maintain or change program activities Maintain or change measures	**Action Component** Decisions concerning Objectives Program elements Monitoring

Source: Adapted from Wholey (1979, p. 118) and Poister (1983, p. 11).

Table 3
R & D Approach to Performance Monitoring

Performance Monitoring Component	Processes	Research & Development Activity
Establish objectives	Develop system objectives with stakeholders	Assessment of stakeholder group satisfaction
Select indicators	Develop criteria selection & measures with stakeholders;	Assessment of stakeholder group satisfaction
	Tested availability of data and quality of data	Pilot data collection and evaluation Validated alternative measures of SES
Comparing program performance	Develop alternative measures of SES; Develop fair comparison methods	Tested "benchmark" method for fair comparisons
Using information	Reporting data	Tested alternative display formats

feedback on the program: teachers, district superintendents, school board members, and representatives of education-oriented organizations (e.g., principals, NAACP, community colleges).

The research and development team, consisting of agency staff and university consultants, conducted periodic, day-long meetings with small groups of representatives of these groups over a two-year period. Selected because of their extensive knowledge of the education system and their high stake in its outcome, these groups guided numerous choices affecting the program components. Stakeholder group participants viewed the overall process and their contribution very positively and appeared to be positive about the program, even though they stand to be evaluated by it (Henry, Dickey, and Areson, 1991). For a description of the assessment of stakeholder group involvement, see the grey box on this page below.

Establishing Objectives

The first step in developing performance monitoring systems was establishing the objectives that will provide a framework for the system. Developing objectives for education is quite difficult because of the myriad societal and individual aspirations that are held for education. Student acquisition of basic skills, values and ethics, socialization, successful performance in the workplace, reduction of welfare rolls, individual self-esteem, proficiency in problem solving and appreciation of the fine arts are among the many goals espoused for education. The objectives selected must be meaningful and important to those who will use the performance data.

The developmental process began with presenting the 14 goals for public education in Virginia that had been established by the Board of Education. In addition, 3 of the criteria for designing indicator systems listed by the State Accountability Study Group were relevant to establishing objectives: (1) Measure the essential features of schooling, (2) measure what is actually being taught or considered important for students to know, and (3) focus on the school site (Office of Educational Research and Improvement, 1988b, p. 7). The objectives are a loose and somewhat pragmatic framework for some of the most important objectives of Virginia's public education system (Table 4).

Three objectives dealt with the endpoints of schooling: graduation, preparation for college, and preparation for work. Each of these objectives encompass all students in the school district to avoid the imposition of labeling students as either college bound or vocational. Collectively, these three objectives promote the recognition that all students should be prepared to successfully undertake further education and to successfully enter the work force. Also, each of these objectives contains leading indicators that relate to the achievement of the objective but are observable early in the educational process (i.e., elementary school).

Another three objectives recognized the three dominant levels of schooling: elementary, middle, and secondary. Too often, indicators have concentrated on secondary school students and the earlier years have been underrepresented. These three objectives are needed to support the process of setting objectives at the school site for all students. Also, they allow for early diagnosis of problems, rather than waiting until the end of the 13-year educational pipeline.

The remaining objective focused the need for information concerning a group of students whose unique characteristics may cause them to be overlooked in the other objectives — students with a handicapping condition. Often these students do not participate in standardized tests under standard conditions and their results are not reported. Even when they are included in indicators such as drop-out rates, their impact is diffused because of their small numbers.

The established objectives to be monitored by the education indicator system were comprehensive in that they included the three most readily agreed upon outcomes of the public education system (i.e., graduation, employment, higher education), and they allowed focus on each of the three levels of schooling. They were also general in nature, which will allow for the inclusion of new and better indicators as the research and development process advances. For example, the Board of Education is currently developing a list of competencies for each student. When the instruments to measure these compe-

Assessment of Stakeholder Satisfaction

Research Problem: Literature on evaluation and performance monitoring emphasizes the need to involve those affected by the system and use their input in guiding its development. Such a responsive, participatory process is hypothesized to increase the quality of a performance monitoring system as well as commitment, support, and utilization among stakeholders.

Sample: The population of participants in the work groups representing educators and education policy makers (N = 58) were surveyed.

Data Collection: Stakeholder representatives were surveyed at the conclusion of the two-year process regarding their awareness, understanding, and involvement in the process; the responsiveness of system developers; their ownership, support, and impact on all aspects of the system; and the quality of the system. The response rate was 95 percent (55/58).

Results: Stakeholder group representatives viewed the stakeholder process very positively, but their impact on the development of the system less so. Teachers, who might be expected to resist such a system, were most positive about their commitment to the system, and impact on it. This may be because teachers are not accustomed to confronting policy decisions and have low expectations for affecting them. Superintendents, have the highest political stake in policy, responded least favorably on commitment and impact. The responses indicated overall that most of the preconditions for evaluation utilization—notably, support, understanding, and commitment—were present for those participating in the process.

tencies are developed, they can be readily incorporated into the existing structure.

Stakeholders viewed the process of establishing objectives as one that they had a great influence over. The opinions of teachers are interesting. In this case and all others, the teachers viewed their participation more positively than the other three groups. In reversing the top-down approach of performance monitoring, the expected aversion of the teachers to the system was avoided.

Selecting Measures

Decisions about the performance indicators to be used require a delicate balancing act. On the one hand, more indicators will provide a comprehensive view of program performance and more data that can be utilized for program improvement. On the other, the costs of the system and the accuracy of the data may be questionable when too many indicators are selected. Poister sets the goal succinctly, "the (performance monitoring system) must be geared to the selective acquisition of 'good' data with maximum potential for meaningful utilization" (1983, p. 12). The first decision concerning indicators was to limit them to measures of student performance, achievement, or accomplishment. Because at the current stage of program development, limited information exists about the input or process variables that affect educational performance, resources were concentrated on outcomes.

The selection process began with a comprehensive list of possible outcome indicators. Literature, other states' lists, existing reports, and original ideas from the stakeholder group meetings were used. Each indicator was evaluated using the following series of seven criteria (McMillan & Henry, 1991):

- ◆ Provides a meaningful measure of the attainment of the seven objectives
- ◆ Encourages positive educational practices
- ◆ Avoids overuse of test scores
- ◆ Provides "leading indicators" that identify negative outcomes later in the educational process
- ◆ Identifies effective educational practices
- ◆ Provides technically valid and reliable data
- ◆ Minimizes data collection burden

Often educational indicator systems rely almost exclusively on standardized test scores. However, test scores provide a limited perspective on student accomplishments and may cause other valued aspects of performance to be ignored (Shepard, 1989; Oakes, 1989). The stakeholders indicated a high degree of efficacy and satisfaction with their influence on the selection of measures. Through the pilot data collection, more issues concerning the last three criteria on the list were uncovered.

Piloting the Collection of Selected Measures

Originally, 63 performance indicators were to be field tested for the 1989-90 school year. Three indicators were dropped because of data collection problems. Field testing showed that an additional 11 indicators were unreliable. For example, counting the number of students involved in cocurricular activities contained duplicate counts in many school districts, resulting in rates of participation over 100 percent.

Public review of the indicators produced a useful distinction between indicators. Twelve indicators were felt to yield meaningful results, but using them for accountability might encourage questionable educational practices (violation of the fifth criterion). For example, knowing the rate of student participation in algebra provides information about whether the students are on track for more advanced mathematics in high school. However, if students are pressed to take algebra before they are prepared, their participation may have a negative impact on them. Therefore, these 12 indicators were classified as "for reporting only" and will not be used for accountability.

The field testing has resulted in obtaining 50 indicators, 38 for accountability and 12 for reporting only. The 50 indicators are shown by objective in Table 5.

Alternative Measures of Socio-Economic Status

One received truth in education is that the socio-economic status (SES) of the family influences student outcomes and that

Table 4
Objectives for Education System

Objective	Definition
1. Preparing students for college	Indicating student knowledge and skills that lead to admission and successful performance in post-secondary education.
2. Preparing students for work	Indicating student knowledge and skills needed in the workplace following high school graduation or post-secondary education that result in successful performance in occupations and professions.
3. Increasing the graduation rate	Indicating student knowledge, skills, and experiences that result in successful completion of high school.
4. Increasing special education students' living skills and opportunities	Indicating special education students' knowledge, skills, and experiences that lead to the development of academic, vocational, and social competencies.
5. Educating elementary school students	Indicating student knowledge, skills, and experiences in grades K-5 that lead to the development of academic competencies and the ability to successfully perform in middle school.
6. Educating middle school students	Indicating student knowledge, skills, and experiences in grades 6-8 that lead to the development of academic competencies and the ability to successfully perform in secondary school.
7. Educating secondary school students	Indicating student knowledge, skills, and experiences in grades 9-12 that lead to the development of academic competencies and result in graduation and successful performance in college and in the workplace.

schooling variables have less impact than socio-economic variables. Because not all school districts have students from similar socio-economic backgrounds, comparing outcomes without taking parental education, income, and occupational status into account is perceived to be unfair. Yet, constructing a variable and collecting data that summarizes the levels of education, occupation, and income in a school or school district is difficult.

The most readily available measure of SES is the percentage of students that are eligible for free or reduced-price lunches. This measure has been used as a proxy for SES in other states. However, this measure, because it is based on student self-selection has never had the confidence of educators. Specifically, they have feared that secondary school students would disproportionately avoid the stigma of applying for the program and, therefore, bias the measure. To test the bias of the percentage of students eligible for free or reduced-price lunch as a measure of socio-economic status, a research project was designed and carried out (see the grey box on this page below). Based on the results of the study, eligibility for free or reduced-price lunch is a satisfactory measure of socio-economic status for use in developing fair comparisons of educational outcomes. However, the eligibility variable is a better measure of income than education.

A Benchmarking Method for Comparing Program Performance

The most controversial component of the educational performance monitoring system has been the method used to make performance comparisons or set criteria for performance. Indicators provide data on the "health" of some aspects of the educational system. Without a reference point, it is impossible to interpret the system's condition. For example, it is difficult to know if a 9.0 percent drop-out rate is good or bad without a frame of reference. Reference points can be derived in three ways: absolute standards, comparisons with other units, and comparisons with past performance (i.e., improvement) (Oakes, 1986).

To be accepted and used, a standard must be considered fair. Therefore, a commitment was made to the development of comparisons that would take characteristics of the students and school districts that may affect educational outcomes, but are beyond the control of the school districts, into account. These comparisons with other units are to be combined with comparisons of each district's current performance to its past performance as additional years of data are collected.

Two methods are commonly used to adjust outcome measures for student and school district characteristics: comparing outcomes to those of similar districts and comparing outcomes to predicted outcomes, controlling for the characteristics. Nine states use one of these methods for developing comparisons (Salganik, 1990). The methods used to compare school districts with similar school districts have several principle disadvantages. First, those that group or cluster districts into fixed groups may label the districts and cause a negative perception of districts with students that are poor and parents that are less educated. In addition, the arbitrary cut-off points for dividing the clusters may result in two similar districts being placed in different groups. Finally, for the methods that require the prediction of outcomes, the choice of which equations to use and the accuracy of the prediction can affect the reference points being used for any district.

While the results using either of these two methods are very close (Salganik, 1990), the stakeholder groups were very

Measuring Socio-economic Status at the School Level

Research Problem: Family socio-economic status is a factor beyond the control of the school that affects student achievement. Thus, it should be taken into account in order to make fair comparisons of student performance at the school level. Measuring socio-economic status at the school level is difficult. Alternative measures should be compared to understand their differences.

Population: The population included the entire population of schools in Virginia (N = 1,088). However, subsets of the population were used for the pilot test of the instrument and the analyses. For the principal rating form, 1,023 usable responses were received.

Data: For all schools, the percentage of students eligible for free or reduced price lunches and principal estimates of the socioeconomic status of parents of the students in a school were collected. Principals indicated the percentage of parents within each of six educational, occupational, and income ordinal categories. The rating instrument was pretested with 75 principals and 19 district administrators in five school districts. Other indicators expected to correlate with these measures were also collected, for example, ability scores of entering students, community income, and student reports of parental education.

Methodology: Three tests of the measures were carried out. First, at each level of schooling, SES measures specific to the school were correlated. Second, for schools whose boundaries were congruent with the district boundaries (i.e., only one school of that type in the district), the school SES variables were correlated with community SES variables. Finally, in one large division where the percentage of students whose families received AFDC was collected, that measure was correlated with all school variables.

Results: Eligibility for free and reduced-price lunch was correlated with the SES principal rating at .69 to .43, depending on the level of school. The principal rating has a much higher correlation with student reports of parental education (.49 & .61 vs .32 & .38) than the eligibility variable. For elementary and middle schools the eligibility variable was more highly correlated with income variables than the SES rating, but the relationship was reversed at the secondary level. However, differences were generally less than .10.

concerned with the implications and potential problems of both. A benchmarking strategy was developed that used each of the 133 operating school districts as a "seed" district and selected the 14 most similar districts based on eight background variables (Henry, McTaggart, and McMillan, 1992). This procedure yields a unique comparison group for each district. Performance of each district is described as "above expectations" when the divisions performance is above the 75th percentile, "within expectations" when performance is within the 25th to 75th percentile, and "below expectations" when below the 25th percentile. The empirical test of the approach is presented in the grey box on page 210.

Reporting Educational Performance

Reporting performance data is a vital, yet often minimally investigated part of developing a performance monitoring system. School personnel must correctly interpret the data if it is to be used. The press, parents, and the public must find the data accessible to be useful for public accountability. Few if any studies have been carried out in a systematic fashion to explore the accuracy, speed and comfort with reporting formats. A few studies do exist concerning accuracy and speed of perception with graphics (Simkin and Hastie, 1987; Stock and Behrens, 1991). Tukey (1988), a pioneer in graphical analysis, has indicated the need for more applications and experiments. Reports used to date have been developed relying solely on anecdotal information about report formats, common sense, and the limitations of computer hardware and software.

To learn more about accuracy and preferences for graphical and table data displays, an experiment was designed and carried out (Henry *et al.* 1991). The results are summarized in the grey box on page 211. For multivariate displays, graphical displays were not as accurately used as tables. However, some design changes in the graphics may have made them comparable. For more dense data, univariate displays, graphs, and tables performed similarly, thus giving the analyst an option. The multivariate display was modified significantly and ultimately the table was included as a supplement to the graphical display.

Summary

Performance monitoring is a natural response to public interest in greater accountability for programs funded by tax dollars and on which society depends for improving the general welfare. Education is one program area where significant public concern has caused the development of performance monitoring systems.

Organizations that are developing a performance monitoring system must make many choices in order to implement the system. These choices can impact the viability and utilization of the system over the long haul. The choices can be made through a research and development approach to test various alternatives and base decisions on empirical data to

Table 5
50 Educational Performance Indicators by Objective

Preparing Students for College	**Preparing Students for Work**	**Increasing the Graduation Rate**	**Increasing Special Ed. Students' Living Skills and Opp.**	**Educating Elementary School Students**	**Educating Middle School Students**	**Educating Secondary School Students**
Receiving the advanced studies diploma Minority students receiving the advance studies diploma* Taking the SAT SAT scores Taking foreign language (8th graders)* Taking Algebra I (8th graders)* Taking advanced placement-college level courses Advanced Placement Test scores* Upper-quartile 11th grade test scores Upper quartile 8th grade test scores Remedial courses College GPA	Occupationally prepared graduates Basic reading skills acquisition Basic math skills acquisition Completed keyboarding or typing	Literacy passport first time pass rate Dropout rate Minority dropout rate* Attendance Above 25th percentile 4th grade test scores Above 25th percentile 8th grade test scores Over age students in 4th grade Over age students in 8th grade	Attendance Dropout rate Receiving regular or advanced studies diploma Literacy passport first time pass rate Work experience Co-curricular involvement*	Above median 4th grade test scores Attendance Literacy passport first time pass rate Over age students in 4th grade Over age minority students in 4th grade* Physical fitness pass rate	Attendance Taking foreign language* Minority taking foreign language* Taking Algebra* Minority taking Algebra* Upper quartile 8th grade test scores Above median 8th grade test scores Physical fitness pass rate	Upper quartile 11th grade test scores Above median 11th grade test scores Attendance Dropout rate Minority dropout rate* Physical fitness pass rate.

* Indicated for reporting only

Testing Benchmark Method for Fair Comparisons

Research Problem: A vexing problem for performance monitoring systems is making judgments about performance. In many cases, factors that affect performance are beyond the control of the organization. A benchmarking method was developed to make comparisons of program performance. Three criteria critical to the acceptance and utilization of the method for making comparisons were ascertained and tested: credibility, predictability, and equity.

Data: Two types of data were compiled: five school district level achievement variables (e.g., test scores); and 63 community and student variables expected to affect educational outcomes (e.g., percentage students eligible to receive free or reduced-price lunch). Analysis was conducted on 133 operating school districts.

Method: A statistical technique was used to set performance standards by providing individual benchmarks (summaries of the 14 most similar divisions) for each school district. The benchmark groups were selected using a multivariate technique based on the square root of minimum squared Euclidean distances (D).

Results: Three tests of the adequacy of the benchmarking method were applied: credibility, prediction, equity.

Credibility: As an initial check on the credibility of the grouping procedure, four panels of educators were asked to group districts into similar input groupings based on their personal knowledge of them without any additional data. The groups selected intuitively by the panels had a high degree of overlap with the benchmark groups produced by the grouping procedure.

Prediction: Benchmark groups are used to make judgments about the performance of districts on educational outcomes. Thus, the more highly related input variables used to develop benchmarks are with outcomes, the better the groupings. An OLS regression equation was fit for the following outcome indicators, using the fewest possible of the eight input variables that improved the R^2, with the derived R^2 given: dropout rate (.20); grade 4 standardized test score (.69); grade 8 standardized test score (.73); grade 11 standardized test score (.63). Although the relationship with dropout rate was less than desired, the variables as a whole were effective predictors of outcomes.

Equity: The benchmark groups must enable judgments about performance to be made equitably. That is, the benchmark groups will be equitable to the extent that the districts have an equal chance of being judged as performing adequately or excellently. The criteria should be equally challenging for districts in a variety of circumstances.

A test of whether it is more or less difficult for districts which have benchmark groups which are less similar (i.e., larger distances between them) to meet or exceed performance expectations than districts with less distance between them was devised. First, criteria were set for performance expectations. Three groups of districts were chosen for the smallest, and the 20 in the middle. An analysis of variance revealed no significant differences between the three D groups in the number of times performance was considered above expectations or substandard on four outcome indicators.

the extent possible. This approach seems to be an obvious one to use for establishing a new program, just as it is obvious in establishing a new product in industry. Yet, it is tremendously difficult to carry out in the public sector.

External groups, especially business interests, want to see an immediate bottom line. Administrators want to know precisely what they are expected to do and for what they are to be held accountable. Traditional bureaucratic culture (i.e., aversion to risk, premium on stability, short-term outlooks) runs counter to the environment needed to foster R&D, making it difficult to utilize for institutionalization.

The research and development approach requires commitment, patience, and openness to ideas and change, attributes that may require significant organizational change and additional resources. For example, in its recent report, the National Panel on Education Indicators recommended that the National Center for Education Statistics (NCES) take the lead in developing a national indicator and reporting system, including shaping analytic and policy use of indicator information. However, the panel recommended that NCES shed its traditional roles and receive funding similar to any private sector R&D effort (Morgan, 1991, pp. 46-57).

The R&D approach offers the opportunity to test alternatives and develop a program in a way that may improve the chances for success—in this case, utilization of the results to improve education. The R&D approach resulted in choices that would not necessarily have been made in the development of performance monitoring, such as the use of benchmark groups, and eliminated errors that would have been made in using some unreliable indicators and less than optimal report displays. It is an approach that can offer advantages for implementing other performance monitoring systems and for other public programs.

Final Note

As a final note, our experiences in using the R&D approach sheds some considerable light on risk-taking in the public sector. Performance monitoring is a risk. Results may indicate that desirable outcomes are not being produced. Risk is exacerbated by stating upfront that research must be used to develop the system. This approach, honest though it may be, is an affront to the "expertise" approach to program implementation. Quite simply, administrators and public officials must constantly expect and manage change, look at the data, and be willing to withhold judgments until the research is completed.

Testing Alternative Displays for Performance Data

Research Problem: Graphics are widely believed to be useful ways to communicate information such as educational performance data. Little empirical research is available on the effectiveness of alternative graphical displays with their intended audiences (Henry, 1993).

Sample: Five populations involved with the use of educational performance data were identified: school board members, teachers, principals, district superintendents, and journalists. Simple random samples of the first three groups were drawn. The population of district superintendents and education print journalists were used.

Data: Experimental data from questionnaire items in which the accuracy of and preferences for alternative display formats were gathered from the participants. An overall response rate of 50 percent was obtained.

Methodology: Two graphical display formats were evaluated, one providing an overview or summary of multiple performance indicators (multivariate display) and another providing more detailed information on individual indicators (univariate display). For comparison, tables were developed containing the same information as each of the two graphical displays. Each participant completed a test of the accuracy with which comparisons in performance could be made and their preferences.

Results: For the multivariate displays, the comparisons favored the table. On average, participants were able to get one more question right with the table ($x = 5.9$) than with the graph ($x = 4.9$), for an accuracy rate of 84 and 70 percent, respectively (significant at $a = .05$). However, the greater accuracy rate for the table may have been as a result of participants having to analyze the graphic format over two pages as well as estimate differences not explicitly displayed on the graph. There were no significant differences between the five groups of participants on accuracy. A preference for the table was indicated on all three Likert-items (significant at $a = .05$).

For the univariate displays, participants were as accurate in judging results and expressed similar affinity for the graph when compared to the table. Participants using the univariate graph were accurate on 67 percent of the items, compared with 63 percent for the participants using the table, a statistically insignificant difference. There was no significant difference between the formats in terms of preference and no differences between the five groups in terms of accuracy.

Many administrators and educators were clearly uncomfortable with the R&D approach because we openly acknowledged uncertainty and insufficient expertise to anticipate all potential problems and fully implement the system. Too often, administrators and elected officials feel pressured to pose *the* solution to a complex problem. The need to gain public confidence and legislative appropriations are factors that weigh against acknowledging uncertainty. Would the appropriations committee look favorably on an approach that needs to be tested and potentially, altered?

These circumstances can create a certainty trap. Claiming certainty of results on the basis of expertise and experience supports the process of program initiation and appropriation. However, it makes evaluation and improvement a risky business. The trap discourages performance monitoring and making changes that may improve the program's impact on those it was designed to serve. Acknowledging uncertainty and using the R&D approach places the risk on the front-end of the program initiation process—can we get it funded? Risk is inherent in the political-administrative environment in which public programs operate. Shouldering the risk in the initiation and development phases and using the R&D approach holds the distinct possibility of benefits in program outcomes. More empirical evidence using the R&D approach is needed to test the approach.

◆ ◆ ◆

Gary T. Henry is the director of the Center for Urban Policy Research and associate professor of public administration and urban studies at Georgia State University. He is the author of *Practical Sampling* and a forthcoming book, *Graphical Display of Data*. He is a former deputy secretary of education and former deputy superintendent of education for the Commonwealth of Virginia.

Kent C. Dickey is a quantitative analyst for the Virginia Department of Education, where he has been involved in Virginia's development and implementation of a statewide educational performance monitoring system. He coauthored an article that appeared in *Educational Evaluation and Policy Analysis* on using stakeholders in performance monitoring.

References

Benveniste, G., 1985. "The Design of School Accountability Systems." *Educational Evaluation and Policy Analysis*, 7(3), pp. 261-279.

Goldman, J.P., 1990. "Grading Schools Through Report Cards." *The School Administrator*, pp. 26-30.

Henry, G.T., K.C. Dickey, and J.C. Areson, 1991. "Stakeholder Participation in Educational Performance Monitoring Systems." *Educational Evaluation and Policy Analysis*, 13(2), pp. 177-188.

Henry, G.T., 1993. "Using Graphical Displays for Evaluation Data." *Evaluation Review*, vol. 17, no. 1. pp. 60-78.

Henry, G.T., M.J. McTaggart, and J.H. McMillan, 1992. "Establishing Benchmarks for Outcome Indicators: A Statistical Approach to Developing Performance Standards." *Evaluation Review*, vol. 16, no. 2, pp. 131-150.

McMillan, J.H. and G.T. Henry, 1991. "Selecting Outcome Indicators for an Educational Performance Monitoring System." Paper presented at annual meeting of the American Educational Research Association, Chicago.

Morgan, A.D. 1991. *Education Counts: An Indicator System to Monitor the Nation's Educational Health*. Washington: National Center for Education Statistics.

Oakes, J., 1986. "Educational Indicators: A Guide for Policymakers." Santa Monica, CA: Center for Policy Research in Education, Rand Corporation.

__________, 1989. "What Educational Indicators? The Case for Assessing the School Context." *Educational Evaluation and Policy Analysis*, 11, pp. 181-199.

Office of Educational Research and Improvement, 1988. *Creating Responsible and Responsive Accountability Systems*. Washington, DC: U.S. Department of Education.

__________, 1991. *Accountability Options: Most Effective When Combined*. Washington, DC: U.S. Department of Education.

OERISASG, 1988. *Measuring Up: Questions and Answers About State Roles in Educational Accountability*. Washington, DC: U.S. Department of Education.

Poister, T., 1983. *Performance Monitoring*. Lexington, MA: Lexington Books.

Poister, T. and G. Streib, 1989. "Management Tools in Municipal Government: Trends over the Past Decade." *Public Administration Review*, 49, pp. 240-248.

Richards, C.E, and M. Shujaa, 1990. "State-Sponsored School Performance Incentive Plans: A Policy Review." *Educational Considerations*, vol. 17, pp. 47-51.

Salganik, L.H., 1990. "Adjusting Educational Outcome Measures for Student Background: Strategies Used by States and a National Example." National Center for Educational Statistics.

Shepard, L.A., 1989. "Why We Need Better Assessments." *Educational Leadership*, vol. 46, pp. 4-9.

Simkin, D. and R. Hastie, 1987. "An Information-Processing Analysis of Graph Perception. *Journal of the American Statistical Association*, vol. 82, pp. 454-465.

Stock, W.A. and J.T. Behrens, 1991. "Box, Line, and Midgap Plots: Effects of Display Characteristics on the Accuracy and Bias of Estimates of Whisker Lengths." *Journal of Educational Statistics*, vol. 16, no. 1, pp. 1-20.

Timar, T.B., 1989. *Educational Reform: The Need to Redefine State-Local Governance of Schools*. Charleston, WV: Policy and Planning Center, Appalachia Educational Laboratory.

Tukey, J.W., 1988. "Some Graphic and Semigraphic Displays." In W. S. Cleveland, ed., *The Collected Works of John W. Tukey*, pp. 37-62. Pacific Grove, CA: Wadsworth and Brooks.

Wholey, J.S., 1979. *Evaluation: Promise and Performance*. Washington, DC: Urban Institute.

__________, 1991. "Intergovernmental Goal-Setting and Ongoing Evaluation: Improving Government Performance and Credibility Throughout the Federal System." Paper presented at the National Conference of the American Society of Public Administration, Washington, D.C.

Wilson, J.Q., 1989. *Bureaucracy: What Government Agencies Do and Why They Do It*. New York: Basic Books.

Wohlstetter, P., 1991. "Accountability Mechanisms for State Education Reform: Some Organizational Alternatives. *Educational Evaluation and Policy Analysis*, vol. 13, pp. 31-48.