Measuring School Improvement: A Few Experientially Based Words of Caution

Susan J. Scollay and Susan Toft Everson

This paper describes the concerns of the authors regarding the use of standardized achievement scores to measure the quality of practice-oriented school improvement activities in school districts. The authors organize their comments into four issues which are related to current educational practice. While the paper focuses on warnings against the misuse of measurement processes, the authors agree with the call for accountability in the education of students. At a time when educators are under pressure to show results, particularly in urban settings, the authors recommend caution in relying exclusively on quantitative measures in judging the complex process of school improvement.

The Mid-continent Regional Educational Laboratory offers an Effective School Program (McREL-ESP) to teachers, building administrators, and central office personnel. Grounded in the research on effective schools and effective classrooms, the McREL-ESP is an intensive, long-term staff development program designed to improve educational practices at the building level. As such, its ultimate goal is to facilitate the improvement of student learning. Our purpose here is to raise a series of cautions based on our experience with the McREL-ESP and within the context of the current national fervor for school improvement. The McREL-ESP has been described elsewhere (Everson et al., 1984). The focus of our cautions is the temptation to use student learning outcomes as a single means of measuring the impact of practice-oriented improvement efforts which are not primarily designed as research inquiries. The discussion covers four interrelated issues.

MULTIPLE FUNCTIONS OF SCHOOLING

First, on a somewhat philosophical level, it is important to remember that much more is learned in schools and accomplished through schooling than can be charted by changes in student test results. Though the current

Susan J. Scollay, University of Kentucky; Susan Toft Everson, Mid-continent Regional Educational Laboratory.

national fervor encourages us to think of student learning in particular and schooling in general in quantitative terms, such an approach can be detrimental to both students and schools in the long run. It is important, therefore, to clearly limit and specify which student outcomes are to be altered by any given intervention effort, and to pick appropriate measures of those outcomes carefully. In that selection process, it is equally important not to neglect some of the less easily measured areas of student growth in the desire to document "school improvement."[1]

PROGRAMMATIC CAUTIONS

On a more pragmatic level our second caution concerns data collection and documentation of whatever specific targets have been selected as the focus of the information. Most school development programs are designed to be practical and action-oriented. This usually means they are not intended to serve the purposes of research. Understandably, few districts have the time and resources required to perform extensive and systematic baseline, formative, and summative data collection and documentation which is necessary for a research inquiry. This does not mean, however, that documentation and data collection should be neglected in development activities. Without baseline data, for example, it is rather difficult to document that desired changes have indeed occurred, much less establish any reliable link between them and any development activities undertaken. If school boards and other constituencies of districts desire evidence of impact and change resulting from development initiatives, school personnel must build in documentation and data collection costs and procedures from the very beginning.

REAL WORLD COMPLEXITIES

Our third and most fundamental caution, however, stems from the inherent complexities of schools operating in the real world. There are some basic realities of planned change efforts operating in real world settings which mediate against establishing a direct cause and effect link between school development efforts and student learning even under the best of circumstances. For example, assume for a moment that a school district undertakes a major staff development effort and has the resources and inclination to support the data collection and documentation noted above. Even in that situation, the complexities of reality argue against the ability to show direct and strong relationships between the professional development activities and changes in student performance in a single building or district. Two graphics adapted from the McREL-ESP documentation and evaluation design will help explicate our point.

Figure 1 outlines the tracing pattern of the intended impact of a typical staff development program for building-level school personnel. As Figure 1 suggests, the ultimate goal of such an effort usually relates to student achievement, and it is to be accomplished by influencing school personnel

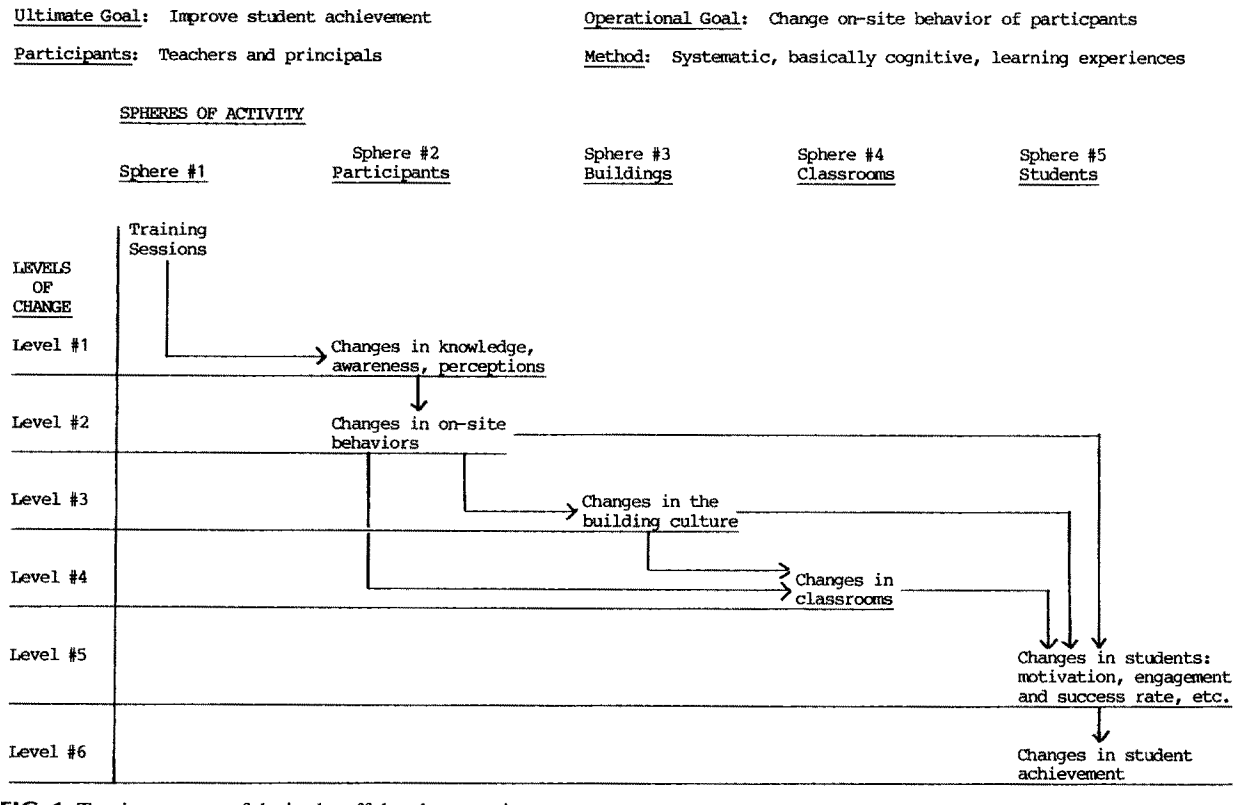

FIG. 1. Tracing pattern of desired staff development impact.

This figure graphically displays the progression of assumptions made in the practice of using staff development to improve student academic performance. To whit, we assume information and insights are transferred from a trainer (and his/her materials) to participants (school level educators) in staff development programs. In turn, we assume this transference stimulates changes in the educator's knowledge, awareness, and perception, which, in turn, will lead to changes in in-school behaviors. By involving several building personnel in this process, we assume that the group effort will support growth in each individual's knowledge, awareness, perception, and behavior, which will then influence and alter the culture of the building. These school alterations will combine with all other influences to affect what happens in classrooms. What we really intend, expect, and assume is that all the changes noted thus far move in a particular and positive direction and that these improvements will stimulate improved student attitudes, motivation level, and attention, which will eventually be reflected in increased student learning.

cognitively and behaviorally through systematic training. In the best of all possible worlds, the content of the development program will be transferred to the participants through systematic means, e.g., presentations, demonstrations, training materials, and/or practice sessions. That transfer will stimulate alterations in participant knowledge, awareness, understanding and/or perceptions, which in turn will be reflected in changed behavior and practice on-site in buildings and classrooms. These latter changes will stimulate alterations in students, e.g., in their motivation, attention, attendance, engagement rates, and ultimately, changes for the better in student achievement. Underlying this multilevel and multisphere desired impact pattern are several characteristics of the practice-oriented staff development program which are realities. These characteristics are both commonsensical and taken as implicit "givens" in the minds of most. They are important enough to make explicit, however:

1. Typically, such a development program is an activist, not a research, endeavor (the purpose is practical results, not data analysis and inquiry).
2. As such, the ultimate goal of the program is to bring about change in targets (student learning and student achievement) which the program does not touch directly.
3. An operational goal of the program is to bring about positive change in a target, on-site behavior of school personnel, which it seldom if ever touches directly.
4. Therefore, to achieve the desired impact of the development program, changes were required on several levels of reality and in multiple spheres of activity before reaching the targets of either operational or ultimate concern.

These realities lead to the second aspect of our most fundamental concern. Figure 2 presents the dispersion pattern of the potential impact of training-based development. Here we have the "flip side" of the desired impact pattern. In Figure 2, it is recognized that even the best practitioners of school development cannot assume that all which is presented is understood, accepted, and internalized. Equally important is the recognition that the environment of development activities is rarely the same as that of the daily context of the school. And most basically, because the content and practice of school development are several levels of change and several spheres of activity away from the target of concern, their impact—both potential and real—is weakened through dispersion before having a chance to reach that target.

As is Figure 1, the impact dispersion pattern outlined in Figure 2 is based on certain realities we recognize but which in the desire or need to show results ignore. These realities include:

1. Even under optimal circumstances, significantly less than 100% of the content of a staff development program is understood, incorporated, and used by program participants.
2. At any given time, a staff development program is only one of a myriad

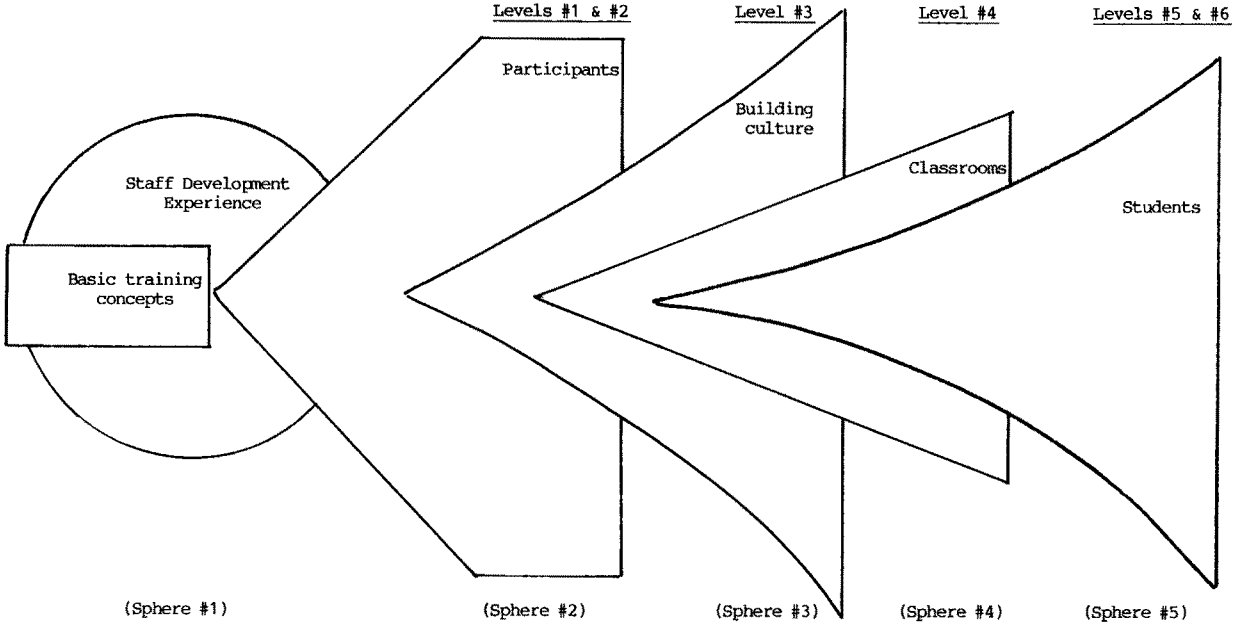

FIG. 2. Dispersion pattern of potential staff development impact.

Figure 2 interjects a sense of reality into the ideal—if complex— situation suggested in Figure 1. Here is a graphic depiction of how the power of any staff development effort is dispersed, and therefore weakened, as it is filtered through the various spheres of activity and levels of change between the experience itself and its intended target, student learning in classrooms. Clearly, less than 100% of the content of a staff development program is absorbed, remembered, or used by the participants. Each participant's knowledge, awareness, perception, and ultimately, behavior is influenced differently, if at all, by what is internalized. Whatever changes do occur may then selectively influence the participant's behavior back in the school, and such changes interact with a complex building culture to stimulate some change in it as well. The process is complex, and there are so many uncontrolled forces influencing what happens, it is difficult, if not impossible, to trace a direct line of causation from the staff development activity through the participant and building culture to whatever changes, and hopefully improvements, may eventually appear in the classroom, in student behaviors, and in student learning.

of factors influencing the knowledge, perceptions, beliefs, behaviors, and daily lives of its participants.

3. Any content transferred from the program to the participants is only one of a myriad of factors determining what happens in the individual buildings and classrooms of those participants.

These graphics and the specifications of limitations they suggest are not presented to discourage or disillusion those who have the desire or necessity to document, assess, and evaluate the impact and results of developmental programs designed to influence the learning experiences of students in school. Such tasks should be undertaken and can be completed successfully.[2]

Equally important, we certainly do not want our cautions to dissuade those who desire to establish some link between the professional development of practitioners and the improvement of student achievement. On the contrary, they are offered precisely because we believe the improvement of student learning is the raison d'etre of educators and that it can be accomplished through systematic professional growth experiences for building-level personnel. Given that, however, we also believe it is important to protect the potential of that process. Thus, it is critical that all involved with such efforts be realistic and candid about what can be achieved, and of that, which portions may and *may not* lend themselves to quantitative measurement and statistical proof.

It should be clear from Figures 1 and 2 that the potential is quite limited for reducing without significant distortion a very complex reality to simple numerical relationships. There are, of course, sophisticated procedures with which this potential may be expanded within the context of a systematic inquiry (Rubin, Stuck, and Revicki, 1982; Fortune and Hutson, 1984). For the action- and practitioner-oriented development program, however, intervening into the real world of schools means operating within an essentially uncontrollable context, one in which accurate statements of direct cause and effect are at best very difficult to make.

ETHICAL CAUTIONS

Our fourth and final word of caution concerns ethics. Recognizing the current pressure on all educators to show positive gains in student achievement, we suggest that the various temptations inherent in that pressure need to be acknowledged. Schooling is a long-term endeavor, and efforts to improve the process can be no less so if they are to be effective. This means the pressure for immediate results must be diffused through enlightenment of the "pressure producers" if long-term and permanent improvement is to be supported and achieved. The process of educating the "pressure producers," usually the public but sometimes the boards of education, is also a long-term one as well, however. Thus it is realistic to assume educators must respond to some of that pressure. If there is an attempt to improve student learning through staff development and to measure that effort in

quantitative terms, at least two basic responsibilities fall to the educators involved.

The first is the responsibility to deal fairly and completely with the data on student achievement. The data should be disaggregated by student groups and analyzed in sufficient detail that any differential effect of the impact claimed can be discovered and assessed.

For example, most urban districts have high student mobility rates, yet they rarely desegregate student achievement data into stable and mobile student groups. In instances of great mobility, to use composite results to measure school progress is a questionable use of such data. Several urban districts in which we have worked rank order schools based on composite student achievement test scores. In one, when the schools with the poorest scores were studied, investigators discovered average student mobility rates of more than 50% in each of those buildings. One elementary building graduated 90 sixth graders, only six of whom started first grade in that school. Even with the most sophisticated quantity/qualitative assessment system, it is difficult to accurately judge school impact when the achievement test population is made up of better than 50% transient students. To make such judgements regarding progress on composite achievement scores causes concern regarding the accuracy of that judgement. An alternative process disaggregating all data into mobile and stable student groups might in fact produce a very different picture of progress. One scenario might include a school where the students who remain in the school for two or more years show gains on achievement tests while the students who come and go test poorly. Obviously, in such a case, the greater the percentage of transient students the greater the negative impact on composite test scores. Rather, disaggregting the data into the stable and mobile groups produces clearer and more accurate pictures of student progress in the school.

Data disaggregation essentially suggests a basic question of equity: Is education improved if only certain segments of the student population experience significant achievement gains? An initial step toward the discovery of any differential effects is to disaggregate the data at the building level. This process allows each school to assess and target its own progress, weaknesses, and strengths. Another important step is to analyze the building-level data and district-level results by variables clearly known to contribute to differential student achievement, e.g., student socioeconomic status or mother's educational background. Obviously, given the sensitivity of this issue—and often the lack of data available on all students—this is not an easy process. But it can be done. There are sufficiently accurate indicators of socioeconomic status on which buildings and districts do have information, (e.g., participation in subsidized school lunch programs), and there are straightforward procedures for analyzing achievement test results by such a variable (Riley, 1983).

The second responsibility incumbent upon educators using standardized achievement test results as a measure of school improvement efforts is closely related to the first and is perhaps even more sensitive. It concerns the dissemination and other uses of the data analyses. Clearly, accountability is

a major issue in both dissemination and other uses of the data. We would not argue against it or against the appropriate use of student test scores. The "hows" and "why" of accountability are, however, as complicated as the process of education to which they are increasingly applied. There is no general "rule of thumb" or packaged program which will work in every district. Particularly in large, urban districts, the politics of the district and its community, the presence/absence of staff bargaining units, and the general morale of school personnel must be considered.

The McREL-ESP has worked with some urban—and rural— districts, which clearly benefitted by wide dissemination of test results and their analyses: these are usually districts which enjoy the basic confidence of their communities and which have a rather strong self-image and sense of competence. In other McREL-ESP districts, however, dissemination of test results, even when they indicated progress and improvement, proved counterproductive because the community was not satisfied, or school personnel were defensive, or both.

A similar situation exists relative to other uses of test score analyses within the context of accountability. Obviously building-level personnel need to know how they are doing relative to past performance, in light of development efforts and, in a general sense, relative to other buildings in the district. To a certain extent, incentives based on quantitatively measured improvements can have a positive effect. Given the complex reality indicated in Figures 1 and 2, with all the potential pitfalls for inappropriate cause and effect statements, we would caution against using student achievement scores as a single, absolute measure. In particular, we caution against the use of such scores in any district competition among buildings or as *the* criterion for evaluation of teachers and/or administrators.

A final word must include a plea for accountability. We would be remiss if readers interpreted our concerns as an argument against the current trend toward educational accountability. Instead, our point is that accountability must be based on appropriate measures, and that both the quantitatively and qualitatively measured outcomes of development work and school improvement efforts be reviewed and carefully reported, including an honest appraisal of the limitations and pitfalls of such endeavors.

The world of schooling is complex; the measurement of schooling and of efforts to improve it therefore is a complex task. In our view, it is overly simplistic to take student achievement scores as the sole criterion for evaluating either the impact of schooling or the quality of school improvement activities.

NOTES

1. It is not within the scope of this paper to discuss all the obvious aspects of student growth which are important to consider, document, and assess, nor can we take the space here to describe how this can be done. We recommend Bodgan and Biklen (1982) as a good reference.
2. Our purpose here, however, is to highlight and warn against the temptation not to do them . . . or to oversimplify process in an effort to document progress or evaluate improvement.

REFERENCES

Bodgan, R. G., and Biklen, S. K. (1982). *Qualitative Research for Education: An Introduction to Theory and Methods*. Boston: Allyn and Bacon.

Everson, S. T., and Scollay, S. J., Vizara-Kessler, B., and Garcia, M. (1984). Application of the research on instructionally effective schools and classrooms: a study of an effective schools project impact at district, school, teacher, and student levels. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Fortune, J. C., and Hutson, B. (1984). Selecting models for measuring change when true experimental conditions do not exist. *Journal of Educational Research* 77(4): 197–206.

Riley, A. (1983). Determining the equity of the effectiveness of a school for different socioeconomic groups (Research on Effective Schools Program, Working Paper 1). Midcontinent Regional Educational Laboratory.

Rubin, R., Stuck, G., and Revicki, D. (1982). A model for assessing the degree of implementation in field-based educational programs. *Educational Evaluation and Policy Analysis* 4(2): 189–196.