ASSESSING THE EFFECTS OF DISTRICT TESTING AND EVALUATION EFFORTS

Donna Mitroff

CSE Report No. 186
1982

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

INTRODUCTION

Over the past two years we have intensely studied six school districts which were identified as involved in activities to link their testing and/or evaluation activities with instruction (T, E, and I). We documented what they were doing, how they came to be doing what they are doing, how they were set up operationally, and how they thought all of these efforts linked testing with instruction. From the outset of the study, we anticipated that stage of the research which asks, in effect, "So what!"

The "so what" question deals with the impact of the testing and evaluation activities of a given district. In the original project proposal, we expressed the issue as that of understanding "the impact or effect of district-wide testing and evaluation activities on the actions of teachers and principals in classrooms and schools." (CSE Plan, 1979, p.18.) We wanted to assess the extent to which the TEI linkage subsystem was having the "desired" or "expected" effects in classrooms.

We have spent considerable time during the past year examining the topic of T/E impact assessment. Through dialogue, review of related work, and some pilot applications we have reformulated the original issue, refined our definitions of T/E impact, and outlined a procedure for others to use for themselves in clarifying their thinking about T/E impact. The purpose of this paper is to share our progress in working through these fundamental methodological issues.

Addressing the Issue

Our original statement of the issue, by focusing on "desired or expected effects in classrooms," suggested that we would look directly in classrooms for teacher behaviors which would indicate that they are selecting and performing actions based on input from testing and/or evaluation data. The first step we should take in designing our research, then, would be to spell out what we would look for as evidence of such effects. In order to do this, we examined what districts were doing in testing and evaluation and attempted to develop reasonable scenarios for what the common impact might be. We asked ourselves: "Given this district's particular testing and evaluation subsystem, what types of effects would flow from it and be evident in classrooms?"

The first thing we learned was that we were confused by the terms we had been taking as synonymous. "Effects," "Impact," and "Use" are not terms that we could continue to use interchangeably unless we wanted to stay hopelessly confused. To clarify our purposes, we adopted the definitions suggested by Smith (1981) for the terms "use" and "impact." Smith defined "use" as "conscious employment of an evaluation (or test) to achieve some desired end or impact," and "impact" as "any discernible actions, events, or changes in conditions that are directly influenced by the evaluation (or testing activities), its processes, products, or findings." Extending these definitions, we add that uses are intended effects whereas impacts can be either intended or unintended. That is, examples of either "uses" or "impacts" can both be referred to as "effects" of a T/E/I linking subsystem.

Returning to the effort to specify indicators, we built a matrix as in Table 1.

TABLE 1: Specifying Effects

	Examples of			
	Uses	Impacts		Sample Indicators
		Intended	Unintended	
Tests				(a) All GATE candidates have 98% CTBS scores
1 - CTBS 2 3 · · · n	(a) Scores will be used to identify candidates for the GATE Program	(c) Teachers identify potential candidates for GATE, notify special services.	(b) Teachers use results to assign reading groups	(b) Examination of classroom reading group profiles shows homogeneous grouping by CTBS scores. (c) Check in cum. cards indicates GATE children who were cited by teachers.
Evaluations				
1 2 3 · · · n				

Using the data from our field work in six districts to try out this matrix led us to the delineation of several key dimensions concerning the assessment of TE Use/Impact. They are listed here in random order.

- Testing and evaluation use and impact can both occur at many levels of the district's organization in addition to the classroom.
- Testing and evaluation use and impact may be understood differently at different levels of the organization, e.g., administrators focus on test scores in relation to an entire program while teachers see scores in relation to individual children.
- Among the variety of observable impacts, some may be intended by someone in a decision-making position; some may be unintended.
- Many of the likely T/E effects are not recognized or articulated by members of the organization.
- There may be a lack of consensus among members of the organization over what constitutes "acceptable evidence" that effects are in fact occurring.
- Among the varieties of effects which can be described, some can be categorized as having a directly technical, that is, instructional, emphasis while others reflect a social/interpersonal emphasis which may or may not indirectly effect instruction.
- Among the varieties of effects some are experienced inside the organization while others are felt in the environment outside the organization, e.g., in the community, by the media, etc.
- Neither anticipated use nor anticipated impact are typically built into the organization as clearly as one would expect.

Assessing the intraorganizational and the environmental effects of testing and evaluation in instruction requires a much broader perspective than that which we had first anticipated. Not only must effects be sought in places other than the classroom; not only must effects on instruction include use as well as intended and unintended impacts; not only must the direct effect on instruction be accounted for, but also the indirect

effects mediated through the social/interpersonal processes.

We must therefore step back to the level of school district ideas, policies, and practices. We need to find a way to assess the effects of both the articulated policies and practices of the district related to the use of testing and evaluation for instructional change as well as the unarticulated intentions. Much of the confusion contributing to multi-level assessments of the effects of testing and evaluation stems from the fact that many district attitudes related to testing and evaluation are not stated as a consistent policy position, but are evolutionary, reactive to circumstances (e.g., public outcry at declining scores), or dependent on the preferences of those in key opinion leader positions.

Therefore, defining the effects that testing and evaluation are supposed to have is not a simple matter of asking one or more policy makers or of searching for a written statement of district policy. Ideas, policies, practices, and expectations change as they filter through the organization and through people's perceptions. The effects--that is, the uses and the impacts of testing and evaluation--occur differentially at different levels as this filtering process occurs.

District intentions regarding testing and evaluation, then, are dynamic in the sense that those who serve as transmitters of intentions are also adding to or modifying the original intentions. This modification occurs at all levels of the organization and suggests that the transmitters or agents of intended policy are also at all levels of the system. To put it another way, ideas and policies are defined and implemented by all those who have a stake in them...all such individuals can be called policy "stakeholders."

This stakeholder concept is of central importance not only to those of us who would like to assess the effects of testing and evaluation on instruction but also to those within a district who would like to manage a testing/evaluation/instruction linkage.

Testing and evaluation are activities which should be carried out with reference to (and deference to) stakeholders. Stakeholders are defined as "those claimants inside and outside of the system who have a vested interest in the problem under investigation and its solution" (Mitroff & Mason, 1981). Those working with the stakeholder concept ask: Who is affected; who has an interest; who is in a position to effect adoption of results or execution of decisions; who has expressed opinions; who ought to care about outcomes? The stakeholder concept is related to the previous statement that effects occur at many levels and are defined differently at different levels, and further complicated by the fact that effects occur both inside and outside of the system. Stakeholders, therefore, may be either internal or external to the school district organization.

To summarize our thinking thus far: the strands which come together are these...

- one cannot effectively assess the effects of a district's testing and evaluation activities without an understanding of the goals and intentions of the district;
- district goals and intentions are embedded in the ideas and in the policies and practices of the district;
- some of the ideas and policies are implicit rather than explicit--unarticulated, rather than articulated; practices may be either consistent or inconsistent with the prevailing ideas and policies;
- the effort to make the ideas and policies explicit must involve a wide range of stakeholders at all levels of the organization.

A PROCESS FOR USE/IMPACT CLARIFICATION

Over the past several months, our project staff has been developing and pilot testing a process to elicit a school district's intentions related to the effect of testing and evaluation on instruction. This process uses a structured workshop in which participants from many levels of the school district organization collaborate.

In addition to enabling a school district to ascertain the implicit expectations of stakeholders at many levels within the organization as well as those outside the organization, it is our hope that the process can be a planning device for districts seeking to create a T/E/I linking system.

The workshop format has two principal justifications. First, it allows participation of individuals from many levels who have diverse perspectives, reflecting our belief that knowledge resides at many levels of the system. It is therefore not sufficient to explore use/impact intentions either for planning or for assessment with input from only the members of a testing and evaluation unit. Second, workshops can incorporate procedures which build "ownership" in ideas or policies. We hope that the dynamic participative workshop procedure that we are constructing meets these conditions.

Our workshop requires a minimum of 8, a maximum of 16 people from across levels and functions in the district. It calls for at least 2½ hours and is even more comfortable if conducted in a longer session.

The workshop procedures are adapted from Mason and Mitroff (1981) who have applied their methodology for dealing with "ill-structured

problems"* in both public and private agencies. We have incorporated their procedures into five major steps which are first listed then discussed in detail.

1. Generate examples of effects of testing and evaluation.
2. Determine the importance and certainty of effects of testing and evaluation.
3. Specify acceptable evidence of effects.
4. Select those effects to be measured.
5. Develop instrumentation to measure effects.

1. Generate Examples of Effects

The basic procedure used in this step is brainstorming -- a somewhat structured form of brainstorming known as the nominal group technique. The nominal group technique requires that the group facilitator go from one person to the next in turn, asking each person to contribute one or more ideas to a group list which is being compiled. The use of nominal group process simply insures that each person in the group has an opportunity to contribute at least one item to the list. The process of going nominally from one participant to the next is continued for successive rounds until members of the group begin to pass. When the point is reached that no other items are forthcoming the process is terminated.

The group facilitator begins the process by asking for examples of some effects of testing and evaluation which participants have observed. No effort is made to focus or channel the items at this point.

---

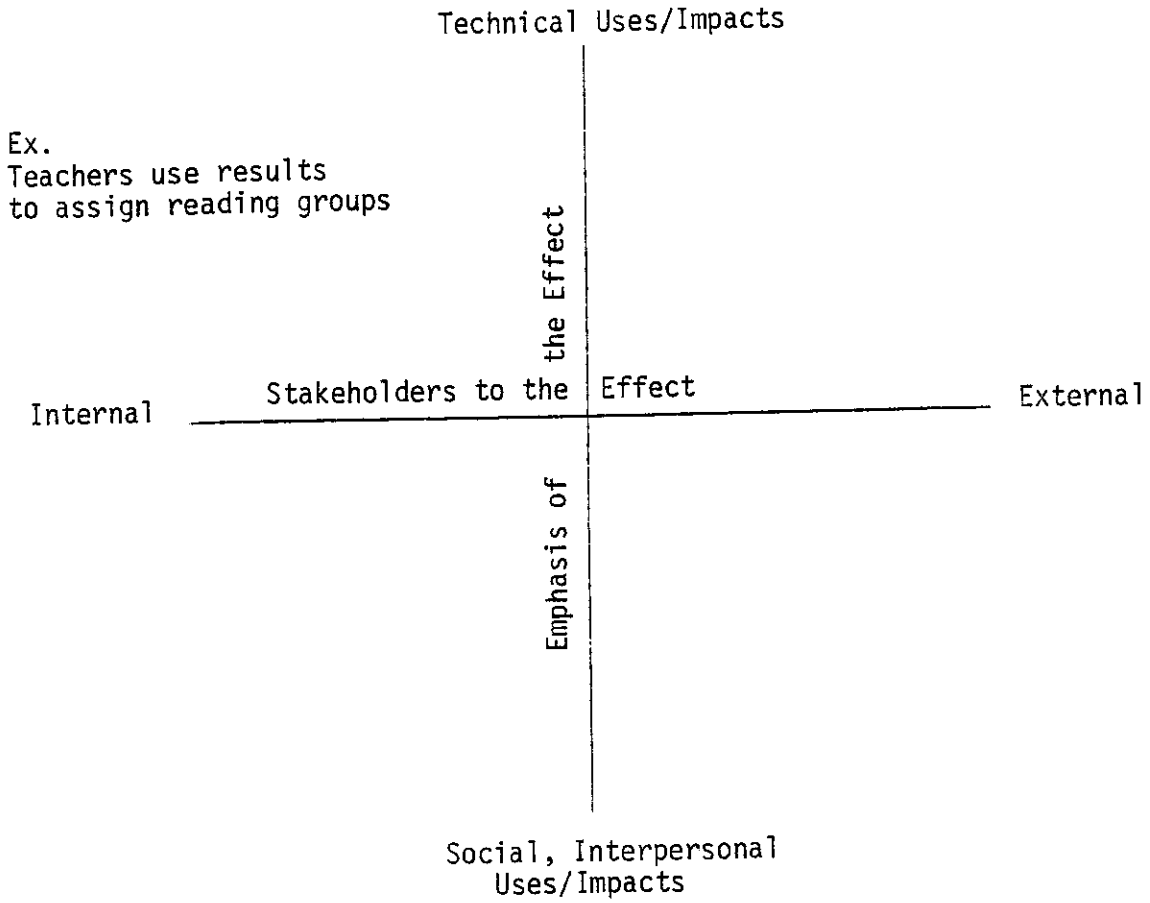
*Ill-structured problems are defined as those for which "there are no single right answers; there is no consensus even on the definition of the problems; and action steps will or should be taken in spite of these ambiguities" (Mason & Mitroff, 1981, p. 29).

Furthermore, no effort is made to judge items contributed to the list.

When it is clear that the brainstorming process has reached an end, that is, when items are no longer coming forth from the group, the facilitator moves on to a sub-step to clarify what has happened in the group and possibly stimulate more thought by clustering the items. A useful clustering strategy can be illustrated by the matrix in Figure 1. We have found that examples generated by these brainstorming sessions can be roughly classified according to two dimensions. The first dimension to be considered is the stakeholders to whom the examples apply and whether those stakeholders are internal or external to the school. The other dimension along which we find a great number of examples clustering is the emphasis of the effect. The emphasis of the effects tend to be either technical or social/interpersonal.

When we take these two dimensions and overlay them to form a matrix, the resulting four quadrants combine different stakeholders with different effects. As an example of the use of the classification scheme, consider one of the examples from Table 1, "Teachers use results to assign reading groups." The stakeholders (classroom teachers and their students) are internal and the effect is technical, i.e., a technical type of instructional decision. This effect has been placed in the upper left quadrant.

Figure 1. Framework for Clustering
T/E Effects



In pilot applications of the process we have drawn this figure on a blackboard or flip-chart and illustrated to workshop participants how their examples can be grouped in one or the other of the quadrants. It is not necessary to classify every item on the diagram -- a representative sampling is sufficient. The facilitator then asks the group if the framework suggests other examples of effects which could be added to the list. Oftentimes it will, and when those are added, the facilitator senses the time to move into step 2.

2. Determine the Importance and Certainty of Effects of Testing and Evaluation

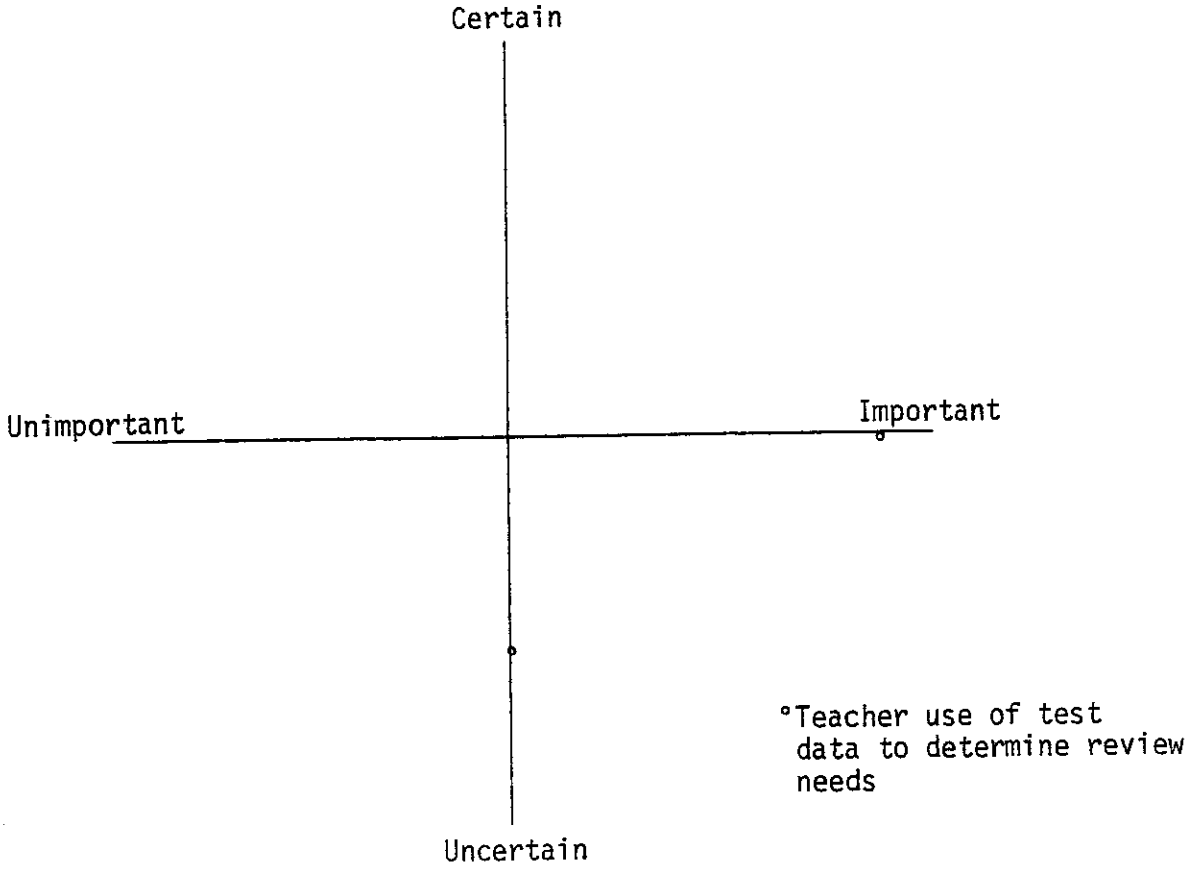
In order to carry out this step the group is broken up into sub-groups. Sub-grouping is helpful because it provides an opportunity for more interaction and more contribution from group members. It also permits participants to consider a selected sub-set of effects rather than the entire range.

To subdivide the group, participants are asked to select themselves, in fairly equal numbers, into one of the four quadrants. One group will then be considering the internal-technical effects, another group will concentrate on the external technical, a third group will consider the internal social-inter personal, a fourth will consider the external social interpersonal.

Once these sub-groups have been formed, the assignment for each group is to take all effects from the brainstorming phase which they feel justified in placing in their quadrant and consider the importance/unimportance, certainty/uncertainty of each of those effects in terms of the district's overall testing and evaluation effort. Once again, participants will be using a cross matrix, as illustrated in Figure 2. This time the horizontal line is scaled from important to unimportant while the vertical line is scaled from certain to uncertain. Each effect is discussed and placed on the classification scheme according to the consensus of the group.

By way of example, consider a sample effect such as the following: test results used by remedial reading teacher to determine consonant blends to be reviewed. The group may agree that such use or impact of test results is important and place it far out on the Important dimension. They may

Figure 2. Determining Importance Certainty



disagree, however, on the certainty of the effect. One person thinks it is a very isolated occurrence; another saying that many teachers use the results in that way. The group decides to place the effect in the important/uncertain quadrant. In so doing they have had "flagged" it as a potential topic for further study and have helped to clarify what it means to them.

Once the sub-groups have completed this step they have, in essence, completed a first sorting of effects and made explicit their notions about which testing and evaluation effects are both "important and uncertain." Effects which fall into that quadrant of the classification scheme are

regarded as the most critical effects according to the judgment of the particular small group. Each sub-group presents its classifications to the entire group. At this point there may be discussion and some re-organizing of priorities. More important though, each effect has been systematically considered.

3. Specify Acceptable Evidence of Effects

Once again, participants work in sub-groups. This time each of the sub-groups is instructed to take those effects which it labeled as "important and uncertain" and produce for each an example of evidence which might be useful in reducing the uncertainty about that particular effect.

For the example used above, the group may decide that interview data indicating that 2/3 of the remedial reading teachers used test results to plan their review activities would be acceptable evidence.

When each of the small groups have completed this task, a general group session is convened for each sub-group to present their list of effects and corresponding examples of evidence.

The purpose of this step is to involve school district personnel in the specification of data sets which they, themselves, will find acceptable. The step is designed to help prevent us, as researchers, from designing and conducting a research study which can be summarily dismissed by its intended clients.

4. Select Effects to be Measured

The effects and evidence lists from each group are compiled into one complete list. It is likely that the composite list is too long for the time and resources of most research. One way to pare it down

is to have workshop participants rank order the total list for research priority. The ranking can be used to determine which effects should become the subject of continued research.

5. Develop Instrumentation to Measure Effects

This step requires that the statements of evidence be used to develop instruments to collect data to prove the presence or absence, strength or weakness of selected effects. In our case the instrumentation is designed by our group, the research team, and presented to the school district team for reaction and revision. Again, the intent is to involve the clients of the research in its design.

The effects clarification process ends at this point. However, the collaborative climate of the process needs to continue through the implementation of the research.

DISCUSSION OF PILOT APPLICATIONS

We have conducted two trials of the clarification procedure. The first was a simulation using members of our Center staff in school district roles. This trial was devised to enable us to try out, revise, and refine the agenda. The second was in one of the sample districts, Northtown, wherein we intended to proceed to develop instruments and assess T/E effects. Recent events in the school district, however, precluded that opportunity and we proceeded only the point of developing a set of research recommendations.

A third trial in another of the sample districts was planned but again events in the district (relevant environments) were such that it could not be carried out.

Staff Simulation

Five staff members assumed the roles of Director of a Testing and Evaluation Unit, District Superintendent, principal, elementary teacher, and secondary teacher. All of the persons assuming roles had, at some time in their careers, worked for a public school system and were familiar with school issues.

In order to set a context for the simulation, participants read a sample district case study and assumed that scenario for their roles.

The trial, abbreviated by the fact that we had only 1 1/2 hours in which to conduct it, was carried out in our conference room. The experience suggested revisions to the process. These revisions will not be elaborated here. Instead, we present the content outcomes of the process.

Table 2. Summary and Classification of Effects from Simulation Pilot

Technical uses/impacts

- Helps teachers make better use of statistics
- Builds curriculum rigidity
- Enables teachers to speak a more common language
- Promotes contact with parents for home management
- Kids are grouped
- Meets individual instructional needs
- Takes up teaching time
- Counselors use to program students
- Provides needs assessment information

- Provides achievement test data for Board
- Accountability function
- Enables school to receive funds
- Provides reports to parents

Internal — Stakeholders — External

Emphasis of effect — to effects

- Kids get grouped for sociometric purposes
- Some kids get upset by tests
- Promotes contact with parents on non-instructional as well as instructional information
- Focuses on cognitive learning; ignores affective domain
- Intensifies competition between teachers and between schools

- Newspaper reports to community
- Gets greater/lesser public support
- Consultation with parents
- Brings recognition from outside sources

Social-interpersonal uses/impacts

Table 3: Listing of Important/Uncertain Effects and Examples of Evidence from Simulation Pilot

Important/Uncertain Effects	Evidence
° Kids are grouped for sociometric purposes	° Examine individual teacher grouping patterns; interview teachers for rationale.
° Focuses attention on cognitive over affective domain	° Observe classroom lessons; examine lesson plans; examine teacher questioning patterns.
° Builds curriculum rigidity	° Interview teachers; examine teacher's objectives over time.
° Enables schools to receive outside funds	° Interview Assistant Superintendent; interview funding sources about why they funded district.
° Brings recognition to district from outside sources	° Interview educational colleagues, applicants for positions in district. Interview researchers at university re district's reputation.

All participants rank ordered the effects listed in Table 3. Their rankings indicated that, were this an actual case rather than a simulation, the research would address the following issues as priorities.

1. Builds curriculum rigidity: (Does the testing and evaluation activity have this effect?)
2. Focuses attention on cognitive over affective domain: (Does the testing and evaluation activity have this effect?)

3. Kids are grouped for sociometric purposes (Do teachers use T/E data in this way?)
4. Enables school to receive outside funds (Have funds become available because of T/E activity?)
5. Brings recognition from outside sources (Can it be shown that the T/E activity has positively increased the district's visibility?)

We point out that while numbers 1,2,3, are classroom based effects, numbers 4 and 5 are the types of effects which can have an indirect effect on classroom instruction.

Because this was a simulation we did not proceed to Step 5 - Develop instrumentation to measure effects. However, the results of the process prepare the research team for that step by providing a client-centered focus.

Participants in the simulation, all experienced with general educational issues and with specific T/E issues, felt that the process brought out aspects of testing and evaluation activities which they had not considered. One of the participants described the process as "a series of sieves through which the issues get refined and focused." They felt the interaction was particularly helpful to their new understandings.

Pilot Application in Northtown.

Through the Assistant Director of the RD and E Unit a two hour session was arranged in Northtown. Discussions with the Assistant Director prior to the workshop determined that the concentration during the workshop

would be on the consolidated application process. That process, defined in another project report (ED Project Annual Report, Nov 1980, p 82) can be described as follows:

- Each consolidated-application school's CTBS scores obtained initially are presented to each school's principal and staff along with the school's mobility index, minority percent, and school enrollment figures;
- Based on these data, the school staff, with the assistance of an Evaluation Services Office evaluator, determines a set of objectives and activities for the coming year. These form the core of the school's annual improvement plan. District evaluators regularly revisit these schools during the ensuing year. The CTBS tests are administered again in the Spring and individual pupil results are reported to the appropriate teacher before the end of the school year. During the Summer, the Evaluation Services Office staff scores the tests and analyzes the results in terms of the individual school's stated goals. A school-specific report is prepared and presented to the school staff in the beginning of the Fall quarter. This forms the basis for the school staff to reformulate goals and activities for the next year -- and the cycle is repeated.

Eight school district representatives participated in the session: there were three from the RD & E unit, two from Title I programs, one principal, and two resource teachers.

We opened the session with a brief summary of the results of our research and an indication of the next phase -- assessing effects. It was clear from the start of the session, and in fact it had been anticipated by the Assistant Director prior to the session, that there were many different agendas on the minds of participants. We also knew that the district was in a state of anticipation of a potential court ruling on desegregation. The many unresolved issues and emotions absorbed some of the allotted time and the full process was not completed.

Table 4: Summary and Clustering of Effects of the Consolidated Application Process in Northtown

Technical uses/impacts

° Report comes out too late to affect activities ° CRT's are used by teachers ° Low CTBS scores led to strict classroom interruption rules ° Process has no bearing on what teachers do in classrooms	° NRT results used by SEA ° RD & E staff conducts in-service when principal cannot. Gets the RD & E staff into some schools ° CTBS analysis led to in-service and special materials ° It satisfies reporting requirements
° Certain minority groups score poorly because of language problems ° Principals set up in-service for PR reasons ° Testing takes up too much student time ° Children are burned out from over testing	° Parents get a better view of what is happening because of report ° Parents in many schools don't understand the report ° Public relations from report is good for some schools - bad for others ° Newspaper publishes NRT results ° Public is fixated on CTBS scores

Internal — Stakeholders to the Effects — External

Emphasis of the Effects

Social-interpersonal uses/ impacts

Beyond the generating of the effects, the process got bogged down and time had run out. When we reflected on what the session told us about the effects of the consolidated application process, we concluded that:

- the process is having its impact mainly in the external technical and external/social-interpersonal area, but very little impact on the internal areas. It would be wrong to even consider effects on classroom activity because such effects have not been built into the consolidated application process;
- the level and form of participation in the consolidated application process is different from school to school and any effort to assess the process needs to use the school site as the unit of analysis;
- the principal's behavior and attitude will be key to level of use in a given school site.

These thoughts along with our suggestions for data collection procedures were shared with the district personnel.

REFERENCES

CSE Three Year Plan proposal to the National Institute of Education. Los Angeles: Center for the Study of Evaluation, University of California, 1979.

CSE Evaluation design project: School district organization study submitted to the National Institute of Education. Los Angeles: Center for the Study of Evaluation, University of California, December, 1980.

Mason, R. O., & Mitroff, I. I. Challenging strategic planning assumptions. New York: John Wiley, 1981.

Smith, N. L. The president's corner. Evaluation News, Summer, 1980, No. 16, p. 24-25.