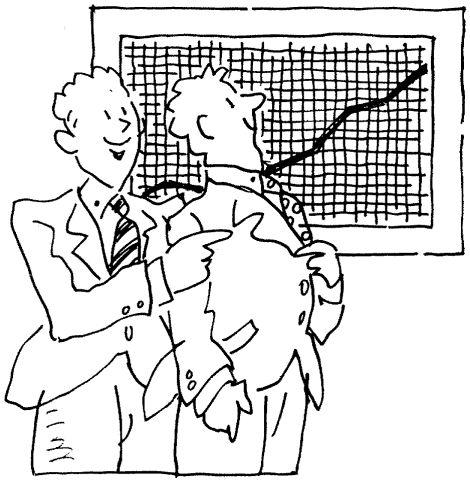
Introductory Packet

Evaluation and Accountability: Getting Credit for All You Do!

(Revised 2015)



*The Center is co-directed by Howard Adelman and Linda Taylor and
operates under the auspice of the School Mental Health Project, Dept. of Psychology, UCLA,
Box 951563, Los Angeles, CA 90095-1563
(310) 825-3634
E-mail contact: Ltaylor@ucla.edu

Evaluation and Accountability:
Getting Credit for All You Do!

Systematic evaluation is increasingly sought to guide operations, to assure legislators and planners that they are proceeding on sound lines, and to make services responsive to their publics.
Lee Cronbach and colleagues

Preface

Evaluation is difficult, and many would prefer to avoid all it entails.

At the same time, most of us agree that schools should be accountable. Yet there are major disagreements about which indicators to collect and how to use the data.

Policy makers demand that schools show that their efforts are effective. But effective in what way? To what degree? At what cost?

In choosing what to look at, how to gather and interpret the data, and what to report, schools reflect society's prevailing values, policies, priorities, and rewards. It is commonplace for there to be dissatisfaction over what is mandated.

Methodologically, evaluation must be carried out with tools that are technically limited. Interpretations of findings are made using different and sometimes biased perspectives. Consequently, what is reported often is controversial.

Moreover, almost everyone has experienced negative consequences from evaluation. Those evaluated often are harmed, and consumers of evaluation reports frequently are misled. Evaluations create tensions and dilemmas and can be misused to create undesirable degrees of uniformity and conformity. Ultimately, we should be as concerned with the consequences of evaluation as we are with improving our evaluation capability.

All this said, perhaps the biggest concern about evaluating and holding schools accountable is that, too often, the processes have not truly improved schools and advanced knowledge.

Evaluation can be the door to a better future. It is more than ironic when prevailing policies and practices close rather than open that door.

Accountability: More than a Mantra

Accountability should not simply be a mantra. It is an invaluable facet of effective practice; but it is just one facet and only makes sense when the other facets are properly planned and implemented.

How effective is the intervention?

Do you have data to support that approach?

Where's your proof?

The questions are so logical and simple to ask, *and* they can be so devastating in their impact. The problem is that such questions imply that relevant data are easy to gather, and so if data aren't available, the intervention must be ineffective or else those in charge are irresponsible. Usually ignored by the questioners are the many complexities associated with valid and ethical evaluation of major mental health and psychosocial problems.

Every educator is aware of the importance of having data on *results.* All interveners want to be accountable for their actions and outcomes. But *it is* complicated.

Fundamental dilemmas stem from the limited validity and focus of available measures and the tendency for those demanding accountability to have inappropriate expectations that there can be rapid improvement even though youngsters and their families are experiencing severe and pervasive problems. Most widely sanctioned evaluation instruments are quite fallible. Moreover, they are designed to measure results that require a lengthy course of intervention, thereby giving short shrift to immediate benefits (benchmarks) that are essential precursors of longer-range improvements. Ironically, demands for accountability tend not to take responsibility for the negative consequences that formal assessment has on some of those who are evaluated. Accountability pressures increasingly require the gathering of a significant amount of data; many professionals note that this practice interferes with building positive relationships and contributes to what is already too high a dropout rate from school and specific helping interventions.

As a result, the topics of evaluation, accountability, and quality improvement are among the most frequent requests for technical assistance and continuing education. The number of publications and technical assistance resources in the area has increased at an exponential rate. And, there are endless lists of measures (many that have not been appropriately validated). Unfortunately, the volume of materials and other resources is not an indication that fundamental evaluation concerns have been effectively addressed. The complications remain unresolved, the status quo remains unsatisfactory; and all that any of us can do at this point is to develop aids, guidelines, and standards for practice that strive for appropriate accountability while doing the least harm to youngsters, their families, schools, special interventions, and society.

As an aid, our intent here is to support evaluative efforts by highlighting a broad range of accountability indicators and outlining ways data related to such indicators currently can be gathered. In doing so, we differentiate three areas for accountability (i.e., accountability to the society, to its institutions that address barriers to learning and teaching, and to youngsters and their families).

Accountability to Who?

In a seminal article on the evaluation of therapeutic outcomes, Strupp and Hadley (1977) stress how different the expectations of society and its institutions often are from those of individual "clients".[1] Thus, it is imperative to understand accountability from the perspective of the various parties with special interests in the results. For our purposes here, the focus is on (a) the *society* in general and the *formal institutions addressing barriers to learning and teaching* in particular and (b) those specific *youngsters and their families* who are the direct focus of intervention efforts.

Accountability to Society and Institutions Providing Student and Learning Supports

Society looks at the following types of general indicators to evaluate whether efforts related to address student learning, behavior, and emotional concerns are paying appropriate dividends:

- *Increases* in youth employment (ages 16-19)
- *Reductions* in
  - >student mobility
  - >youth pregnancy
  - >sexually transmitted diseases
  - >child abuse/neglect
  - >youth arrest/citation
  - >youth probation violations
- *Reductions* in
  - >youth emergency room use for mental health and psychosocial related events
  - >foster care placements
  - >homeless youth
  - >youth suicide rates
  - >youth death rates

In addition, those responsible for schools are required to demonstrate effective fulfillment of their specific mission -- which is to educate the young in ways that meet society's needs. The primary indicators currently demanded by social policy are those that reflect academic achievement at a standard competitive with other major countries. Thus, the emphasis is on increasing

- at all grades
  - >achievement test scores
  - >grades
  - >other indicators of progress in academics (analyses of work)
- at high school level
  - >number graduating (with a related reduction in the number dropping out)
  - >number taking SATs
  - >number continuing with post-secondary education

Because many youngsters are experiencing barriers to learning and performing at school, programs and services to address such barriers are increasingly essential to the ability of schools to accomplish their mission. Some major indicators for accountability related to these enabling or learning support programs are

[1] As these writers state: "*Society* is primarily concerned with the maintenance of social relations, institutions, and prevailing standards of sanctioned conduct. Society and its agents thus tend to define mental health in terms of behavioral stability, predictability, and conformity to the social code. ... The *individual client* ... wishes first and foremost to be happy, to feel content [and] thus defines mental health in terms of highly subjective feelings of well-being ... (p. 188). Strupp, H.H. & Hadley, S.M. (1977). A tripartite model for mental health and therapeutic outcomes with special reference to negative effects in psychotherapy. *American Psychologist, 32*, 187-196.

- *Reductions* in
  >unexcused absences
  >tardies
  >suspensions/expulsions
  >referrals for misbehavior
  >referrals for learning problems

- *Increases* in
  >attendance
  >cooperation & work habits
  >fluency in English as Second Language

- *Reduction* in numbers designated as Learning Disabled, ADHD, or Emotionally Disturbed

Data for Accountability to Society and the Institution of Schooling

Data related to most of the above indicators are available from the records at school sites, school districts, and city/county agencies. Some schools also are involved in administering the Youth Risk Behavior Surveillance System (sponsored by the Centers for Disease Control and Prevention) which contains relevant indicators for use in monitoring changes over time. (Many communities and child advocacy groups are gathering local and statewide data on child well-being and publishing it as "Report Cards.") Various governmental units and other organizations also publish reports of child status indicators. If data are not available, then efforts are needed to ensure relevant indicators are gathered and made accessible. And, appropriate steps should be taken to ensure that data can be disaggregated with respect to specific subgroups.

Accountability to Specific Youngsters and Families

Those who work in school districts to provide special programs and services also are accountable to the specific individuals they help. Such accountability certainly can be seen as encompassing the indicators listed above. However, for individuals who must deal with major barriers, many of the above realistically are only good indicators of progress after a lengthy period of multifaceted, comprehensive, integrated intervention. More immediate accountability indicators are needed to demonstrate progress related to objectives that are the current and direct focus of prevention and special assistance interventions (e.g., reductions in symptoms; enhanced motivation and psychological and physical well-being). Because data on such specific objectives are not readily available, the problem of *generating* relevant data arises -- as do some serious dilemmas. Efforts to answer the following questions lead to an appreciation of the many problems and issues.

What are the right indicators?

Endless arguments arise over indicators when they are discussed in highly *specific* and *concrete* terms. At a more abstract level, there is considerable agreement around three general categories: (1) "client" satisfaction (youngsters; families, schools, society), (2) reduction in youngsters' symptoms/problem behaviors, and (3) increases in positive functioning (youngsters; families, schools, society).

*How can appropriate **specific** and **concrete** indicators be identified for designated "clients"?*

The dilemmas that arise here reflect the problem of "Who is the client?" – students? families? schools? society? all of these? Additional dilemmas arise because the various involved parties

often have different perspectives regarding what problems should be addressed. (And , of course, the intervener may have even another perspective.)

How should the deficiencies associated with existing measures be accounted for?

Although some measures are better than others and some are designated the best that exist, best should not be equated with good or good enough. All instruments we rely on currently have limited reliability and validity; also quite limited are the normative data for various subgroups. These limitations (1) call for care in using any assessment procedure, (2) require full disclosure of limitations when findings are reported, and (3) warrant making extreme efforts to look for disconfirming evidence and interpretations whenever findings suggest significant problems/pathology or positive outcomes that are a bit too self-serving.

How can the negative impact of gathering the data be minimized to an appropriate degree?

All evaluation has the potential to produce major negative consequences. The ethical obligation is to maximize benefits and minimize costs to clients. Putting aside the financial costs, it is clear that use of any formal measure can increase distress for students, families, and school staff and produce psychological reactance.

Sampling of Indicators with Respect to Different Accountability Demands

As should be evident from the preceding discussion, it can be extremely costly and time consuming to be accountable to all parties (see also Figure 1). In most situations, the reality is that only a sample of data can be gathered (see Figure 2).

In working with students, assessment should begin with a focus on data that has direct and immediate relevance to specific intervention objectives. Then, in response to accountability demands and in keeping with ethical and feasible practice, a subset of standardized items can be administered, preferably to stratified samples of clients. The particular subsets of items chosen should reflect matters of greatest concern to those demanding accountability. If the pool of items is large, then different subsets of items can be administered over time and later combined to provide a full picture of outcomes.

With respect to societal and institutional accountability, the data sample initially consists of that which can be readily gathered on a regular basis. Subsequently, again reflecting matters of greatest concern to those demanding accountability, step by step strategies can be developed to establish systems for amassing regular findings related to key variables and specific population subgroups.

Clearly, sampling requires considerable planning and careful implementation. A systematic evaluation plan must be developed, and there must be appropriate budgeting for its implementation. Many programs will require specific consultation in developing an appropriate sampling strategies.

Standards for Comparison

Whatever data are collected will be imperfect and only rarely will be easily interpreted. For accountability to be rational, there must be a reasonable set of standards for comparison. In asking how good an intervention is, the question must be answered in terms of *Compared to what?*

When it comes to student and learning supports, the best comparisons are (a) data on the previous results of intervention efforts with comparable students and their families, (b) data on similar students/families at a school who have not yet been served (e.g., appropriate waiting list samples), or (c) data from a very similar school that does not have the programs being evaluated. The first approach calls for gathering a “baseline” of data before or in the early stages when an intervention is being developed. The latter approaches call for being able to gather the same data with nonserved groups. Again, the matters of systematic planning and appropriate budgeting are central considerations.

Finding out if interventions are any good is a necessity. But in doing so, it is wise to recognize that evaluation is not simply a technical process. Evaluation involves decisions about what, how, and when to measure, and these decisions are based in great part on values and beliefs. As a result, limited knowledge, bias, vested interests, and ethical issues are constantly influencing evaluation processes and the decisions made with respect to accountability.

Figure 1. Accountability as related to different intervention goals.

Purpose of Intervention	***Accountability to Who?***	***Sampling of Objectives/Goals***
To meet *society's goals*	Accountable to society	Data are gathered on indicators that reflect society's purposes in financing the institution
To meet an *institution's goals*	Accountable to a specific institution	Data are gathered on indicators that reflect the institution's purposes
To meet the *personal goals of specific "clients"*	Accountable to specific "clients"	Data are gathered on indicators that reflect individual "client's" purposes in participating in an intervention
To meet some combination of society, institution, and individual goals	Combination of the above	Combination of the above

Accountability: Accounting for Motivational Differences

Pressure to gather accountability data seems like a straight forward practice when viewed through the lens of evaluation. But often those we are trying to help have a negative view of the matter and that produces psychological reactance and negative motivation.

It has always been hard to get the involvement of some youngsters and their families in certain forms of special assistance. For example, no-show and drop out rates are high related to prescribed counseling. The lesson of all this is that greater attention must be paid to enhancing the *motivational readiness* of those we want to help and, at the very least, interveners must minimize doing things that increase avoidance tendencies among clients.

In terms of everyday practice with youngsters, this means (1) discussing with parents/teachers what they should and should not say to youngsters in preparing them for the first visit to student support staff and (2) designing first visits around the concept of enhancing motivational readiness. Everyone needs to be honest and nonpunitive with a youngster in discussing who they are going to see (e.g., a counselor who will try to help make things better). Many students require a great deal of reassurance because of the potential stigma of being sent for special assistance.

Obviously, youngsters who already are well motivated require little to enhance their motivational readiness. For most, however, enhancing motivation toward participation in the intervention requires considerable attention. Because referrals for special assistance usually far exceed resources, it is commonplace for interveners to let difficult, unmotivated clients "drop out." Higher standards of practice call for intensive efforts to enhance the motivation of such individuals so that their problems can be addressed.

The fact of major differences in motivational readiness points to the need to measure such differences so that data on results can be disaggregated with respect to initial motivation and subsequent shifts in motivation. Failure to account for motivation differences reflects serious naivete about the complexities involved in addressing the problems of youth.

Section I B:

Rethinking Evaluation and Accountability

We begin rethinking evaluation and accountability by viewing evaluation and accountability from the perspective of "indicators" as they are used in education. This is followed by an exploration of some specific considerations and concerns that arise in evaluating results and a discussion of the need for policy makers to expand the accountability framework for schools.

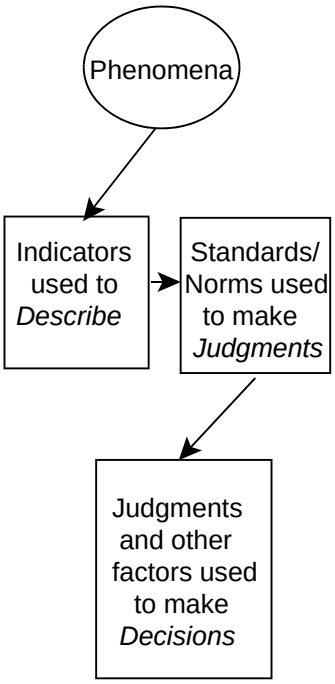
Given the large emphasis that society places on using indicators as measures of performance and descriptions of status and change, the rarity with which we exercise any critical assessment of these measures is surprising.

Planty & Carlson (2010)

With accountability mandates and the emphasis on science-based interventions, the term indicators is widely used in school circles. However, it often is employed so generically that many users are unclear where it broadly fits into assessment and evaluation practices and decision making for policy and planning. We begin our discussion of rethinking evaluation and accountability by viewing these comcepts from the perspective of "indicators" as they are used in education.

What Are Indicators in Education?

Simply stated, an indicator is a pointer that has been developed to focus on and usually quantify useful information about a matter of interest (e.g., current status of students, school performance). Indicators vary in the degree to which they provide direct information, usually expressed quantitatively. Single statistic indicators (e.g., dropout rates) often are used in conveying the status of the education system. Several indicators can be combined into a composite index that encompasses multiple dimensions and increasingly complex phenomena.



Indicators have been described as bits of information that help clarify the characteristics and status of individuals and systems, highlight changes, underscore distance from intended goals, and project the future. To these ends, indicators may focus on inputs, processes, outputs, and/or outcomes.

The importance of indicators stems from reasons underlying their use. Broadly defined, indicators can be used to meet education's many *assessment* needs and play a major role in policy and practice decision making. In general, indicators are used to monitor and describe a myriad of phenomena in order to make judgments ranging from extremely positive to extremely negative. Of particular concern are current conditions relevant to learning and progress in achieving immediate objectives, intermediate goals, and long-term aims. Indicators also are used to project what the future will bring.

Shavelson, McDonnell, and Oakes (1991) proposed the following working definition as a heuristic guide: *An indicator is an individual or composite statistic that relates to a basic construct in education and is useful in a policy context.*

They state:

> "Education indicators are statistics that reflect important aspects of the education system, but not all statistics about education are indicators. Statistics qualify as indicators only if they serve as yardsticks. That is, they must tell a great deal about the entire system by reporting the condition of a few particularly significant features of it. For example, the number of students enrolled in schools is an important fact, but it does little to tell us how well the education system is functioning. However, data on the proportion of secondary students who have successfully completed advanced study in mathematics can provide considerable insight into the health of the system, and can be appropriately considered an indicator. ...
>
> Whether indicators are single or composite statistics, a single indicator can rarely provide useful information about such a complex phenomenon as schooling. Indicator systems are usually designed to generate more and more accurate information about conditions. However, an indicator system is more than just a collection of indicator statistics. Ideally, a system of indicators measures distinct components of the system and also provides information about how the individual components work together to produce the overall effect. In other words, the whole of the information provided by a system of indicators is greater than the sum of its parts."

An indicator is based on underlying assumptions and is only one source for understanding a complex phenomenon

Indicators often are delineated to capture complex phenomena related to students and schools (e.g., student achievement and engagement, school improvement). In such cases, the indicators are focused on *construct*s (i.e., abstract terms used to represent complex concepts). Constructs, of course, are difficult to measure. The construct *engagement* exemplifies the point. This multi-dimensional construct has been divided into three categories – behavioral, emotional, and cognitive (Fredricks, Blumenfeld, & Paris, 2004). Indicators used to measure the construct overlap the three dimensions and are highly correlated with each other. But the three dimension correlate differently with such outcomes as achievement, attendance, and dropping out (National Center for School Engagement, 2006a, b). Other prominent examples of indicators for school-relevant constructs are found in the literature on *school climate* (National School Climate Council, 2010) and *social emotional development* (Isakson, Davidson, Higgins, & Cooper, 2009, 2011).

Indicators & Accountability

Indicators are used as a core facet of accountability and related decision making. Indeed, much of the current emphasis on indicators in education derives from accountability demands (as evidenced by the many formulations of education indicators on the internet).

Thus, the prevailing cry is for specific outcome evidence – usually in terms of readily measured immediate benefits – and for cost containment (see Exhibit 1). Although understandable in light of the unfulfilled promise of so many programs and the insatiable demands on sparse public finances, a narrow results emphasis can be counterproductive. That is, while the prevailing sets of short-term outcome indicators are used as red flags, these indicators alone do not clarify trends or cause and effect, often gloss over important subgroup differences, and rarely include a focus on unintended results. Thus, accountability indicators alone offer too little information to guide practices for improving results.

Exhibit 1

Key Performance Indicators and Leading Indicators

The emphasis on accountability has generated considerable discussion of (1) Key Performance Indicators and (2) Leading Indicators.

(1) As summarized in Wikipedia, "Key Performance Indicators define a set of values used to measure against. These raw sets of values, which are fed to systems in charge of summarizing the information, are called indicators. Indicators identifiable as possible candidates for KPIs can be summarized into the following sub-categories:

>Quantitative indicators which can be presented as a number.
>Practical indicators that interface with existing company processes.
>Directional indicators specifying whether an organization is getting better or not.
>Actionable indicators are sufficiently in an organization's control to effect change.
>Financial indicators used in performance measurement and when looking at an operating index

Key Performance Indicators, in practical terms and for strategic development, are objectives to be targeted that will add the most value to the business. These are also referred to as Key Success Indicators." (http://en.wikipedia.org/wiki/Performance_indicator)

(2) *A leading indicator* is a statistic that predicts trends, usually economic trends. For example, in education, improved average yearly attendance may be a leading indicator for long-term school improvement. However, short-term increases or decreases in a leading indicator often are not predictive of longer-term trends. Besides achievement indicators (including the achievement gap), examples of other leading indicators in education are attendance/truancy, school attachment/engagement, and dropout/graduation rates.

An example of the use of leading indicators for schools comes from Iowa (Iowa Department of Education with the Iowa Collaboration for Youth Development, 2004). The state's results-oriented approach has specified six long-term aims and a set of leading indicators related to each. Note that the indicators include both school and community data:

1) All Iowa youth are successful in school.
Leading Indicators:
% of 8th graders proficient in reading
% of 8th graders proficient in math
Average daily attendance rate
% of students who drop out of school (grades 9-12)
% of youth who are committed to school/learning

(cont.)

2) All Iowa youth are healthy and socially competent.
   Leading Indicators:
   - Rate of juvenile delinquency complaints per 10,000 youth
   - % of youth reporting not using alcohol, tobacco, and other drugs during last 30 days
   - % of youth reporting they have neither planned, considered, nor tried to commit suicide
   - % of youth who report that they have not engaged in violent/aggressive behavior

3) All Iowa youth are prepared for a productive adulthood.
   Leading Indicators:
   - Rate at which students graduate from high school.
   - % of 16 – 19 year olds who are not in school and who are not working.
   - % of youth who report that they help others 3+ hours/wk
   - % of 11th grade youth who report that they work 3+ hours per week in paid job
   - Rate of births to teen aged mothers age13-17

4) All Iowa youth are in safe and supportive schools.
   Leading Indicators:
   - # of long-term suspensions or expulsions for violent crimes on school grounds or at school-sponsored events
   - % of youth who report that staff and students at their school support them.
   - % of youth who report that the norms of the peers in their school are positive.
   - % of youth who report that they feel safe at school.

5) All Iowa youth are in safe and supportive families.
   Leading Indicators:
   - Rate of children found to be neglected or abused
   - % of families in the child welfare system
   - % of youth who report that their families are involved with and support them.
   - % of youth reporting that their families provide them with boundaries

6) All Iowa youth are in safe and supportive communities.
   Leading Indicators:
   - Rate of adult arrests
   - % of families living below the poverty level
   - Rate of persons who are employed
   - % of youth who report that their neighborhoods are safe.
   - % of youth who report that their neighborhoods are supportive.

On a national level, Child Trends (2009) has suggested core goals, desired results, and data sources related to 21 leading indicators for the *Promise Neighborhoods* initiative.

> *Indicators must be related to one another so that their relationships, and changes in these relationships, can be ascertained to suggest possible explanations for observed changes in outcomes.* (Shavelson, McDonnell, & Oakes, 1991)

Criteria for Choosing Indicators

If data are already being gathered using a broad band set of indicators, available data may be sufficient. Otherwise proposals for new indicators must be formulated in ways that enhance rather than detract from an integrated approach to meeting education's many data needs.

In choosing indicators, the emphasis is on considerations such as (1) relevance, (2) how useful and timely the data will be,

(3) how reliably and validly the indicators can be measured, (4) which indicators already are being measured and what it will cost to amass existing data, and (5) what it will cost to gather and analyze data related to new indicators. These considerations are of particular concern when new initiatives and specially funded projects are adopted and call for additional data.

Using Indicators in Schools for More than Accountability

Understanding education's *assessment* needs and practices provides a broad context for thinking about indicators. Formally defined, assessment is the process by which attributes of phenomena are *described* and *judged.* Descriptions take the form of data gathered by formal and informal measures, such as tests, structured observations and interviews, self-reports, surveys, available records, and so forth. Designated indicators guide what is and isn't measured. Judgments take the form of interpretive conclusions about the meaning of data, such as whether a phenomenon is good or bad, above or below standard, dysfunctional or not. Judgments may represent a conclusion about the past, a statement about the present, or a prediction about the future. Judgments inform decision making.

Schools need to pursue assessment related to various functions. Besides system management data, assessment plays a key role in:

1. ***Identification.*** Indicators are used to help find and label phenomena of interest. The focus may be on person variables, environmental factors, or both, and on problems, strengths, or both (e.g., data to inform identification of effective teachers and effective schools; data to inform identification of gifted and talented students and those who are not doing well at school – including those needing special education).

2. ***Selection.*** Indicators are used to help make decisions about general changes in status (e.g., data to inform decisions about moving teachers and principals to different schools, choosing schools for special intervention, placing students in specific programs,).

3. ***Planning for specific changes.*** Indicators are used to decide about immediate and short-term objectives and procedures for accomplishing long-term goals (e.g., data to inform school improvement planning, professional development, specific student interventions – including data from response to intervention efforts and IEP assessments).

4. ***Evaluation of School Results.*** Indicators are used to decide effectiveness based on positive and negative outcomes and related costs (e.g., focus may be on impact on students, particular subgroups, society as a whole). Data are used to make decisions for system improvement and policy purposes (e.g., accountability).

From this perspective, identifying or formulating indicators begins with clarity about functional needs. And, clearly, most of the above functions call for more than indicators of results.

Indicators Raise Methodological, Political, and Policy Concerns

All assessment in education is a complex matter, and controversy surrounds prevailing approaches. Some of the controversy is about the deficiencies and limitations of specific procedures (e.g., lack of standardization, poor reliability, poor validity). Broader concerns have been directed at the way assessment is practiced in schools (e.g., an overreliance on indicators of results often means that antecedent conditions/inputs and transactions are given short shrift. Political and policy concerns have been raised related to the way overreliance on indicators of results has reshaped what schools do and do not do (e.g., Ravitch, 2010).

It is important to remember that choices about what data to gather and exclude are guided by policy decisions, and major decisions about education involve considerations that go well beyond the availability of valid data. Profound and conflicting social-political-economic-philosophic agenda are at play; so no one should be surprised that relevant data often are ignored, and some data are manipulated during policy debates and at decision making tables. As Rutkowski (2008) cautions, "Through educational indicators a set of 'truths' is arguably produced. However, these 'truths' are very open to interpretation." And as Planty and Carlson (2010) stress "Indicators of poor quality certainly distort and misguide decision making and policy."

Some Specific Considerations and Concerns About Evaluation of Results

Two unfounded presumptions are at the core of most current formal and informal evaluations in education. One premise is that an intervention in widespread use must be at a relatively evolved stage of development and, therefore, warrants the cost of summative evaluation. The other supposition is that major conceptual and methodological problems associated with evaluating intervention are resolved. The truth is that interventions are frequently introduced prior to adequate development, with a view to evolving them based on what is learned each day. This is even true of many empirically supported practices brought to schools. (Remember: efficacy

data does not predict effectiveness when implemented by school personnel under common school conditions.)

Moreover, many well-institutionalized approaches remain relatively underfunded and underdeveloped. As to the process of evaluation, every review of the literature outlines major unresolved concerns. Given this state of affairs, the nature and scope of accountability demands often are unreasonable and chronically reflect a naive view of research and theory.

Evaluation involves determining the worth or value of something. As an assessment function, evaluation is defined as a systematic process designed to describe and judge the overall impact and value of an intervention for purposes of making decisions and advancing science.

Properly developed, a set of evaluative indicators can aid efforts to (1) assess efficiency, effectiveness, costs, and impact, (2) make decisions about what to do to improve schools, and (3) advance knowledge in ways that can enhance understanding of and improve policy, practice, training, and theory.

Not all indicators are created equal.
Planty & Carlson

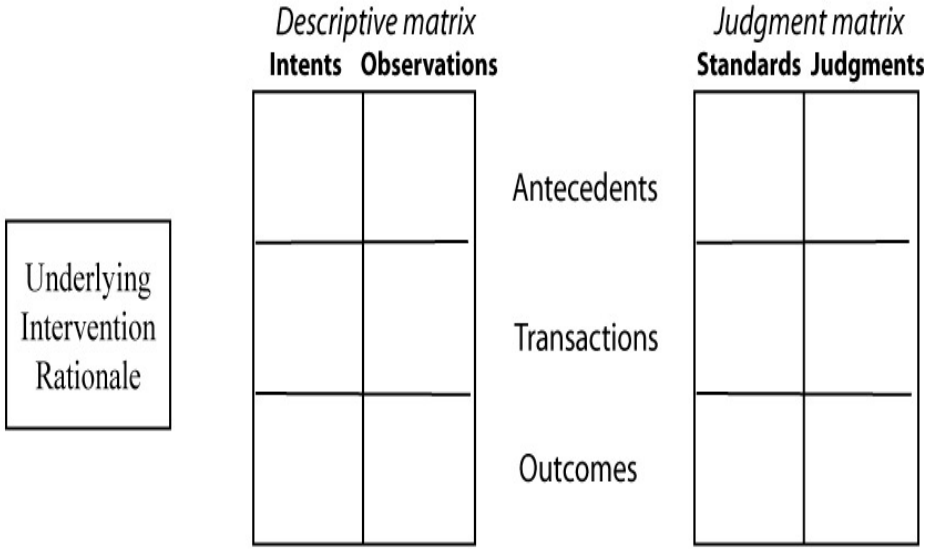
Given that many more indicators can be formulated than can be feasiblely used, decisions must be made about what will be evaluated. In addition to matters highlighted above, these include decisions about (1) the general phenomena of interest (e.g., students, teachers, support staff, administrators; classroom and schoolwide conditions and climate; intervention antecedents/inputs, immediate objectives, intermediate goals, long-range aims), (2) the specific facets to be evaluated, (3) the level of specificity used in designating indicators, (4) the measures and methods for gathering data on designated indicators, and (5) the standards to be used in analyzing the data and arriving at judgments. In making such decisions, concerns arise because what can be evaluated currently is far less than what schools state as their mission. Furthermore, all such decisions are influenced by various sources of bias.

A model formulated by Robert Stake illustrates the type of frameworks used to clarify factors influencing outcomes (see Exhibit 2). Stake stresses that program evaluation requires data and criteria for analyzing the degree to which

- conditions anticipated prior to the program (antecedents), planned procedures (transactions), and intended outcomes are consistent with the program rationale and are logical in relation to each other
- intended antecedents, transactions, and outcomes actually occur.

Exhibit 2

A Framework for Program Evaluation



Source: R. Stake (1967). The countenance of educational evaluation. *Teachers College Record, 68*, 523–40.

In general, the types of data Stake's framework calls for can provide a wealth of information for use in describing and judging school improvement efforts and making decisions about ways to enhance such efforts. Clearly, the data can be used for purposes of accountability, but also for guiding improvements and building an empirical body of effectiveness data.

Note that evaluations of whether a practice or set of practices is any good must first address the question: *Is what is to be accomplished appropriate?* The frame of reference for such evaluations may be the underlying rationale or what others think the practices should accomplish or both. After judging the appropriateness of what is wanted or expected, the intended breadth of focus should guide efforts to evaluate effectiveness. Because not everything is measurable in a technically sophisticated way, some things will be poorly measured or simply reviewed informally. Obviously, this is less than satisfactory. Still, from a rational perspective, continued emphasis on the entire gamut of what is intended is better than limiting things to what can be measured readily or to naive accountability demands.

Finally, we stress that evaluative practice can produce negative effects. For instance, over time, what is evaluated can inappropriately reduce and reshape what a school does and doesn't do. The process is especially pernicious when indicators are used to oversimplify the complex nature, scope, and aims of education.

A Categorization & Examples of Short-term Outcome Indicators for School Use

Efforts to categorize and provide specific indicators for schools generally are concerned with both academic learning outcomes and practices identified as contributing to such outcomes. A prominent example of the latter are the categories and specific indicators for effective school practice developed by the Center on Innovation and Improvement (http://www.centerii.org/handbook/Resources/Appendix_Indicators_school.pdf).

Analyses of categories guiding evaluation of school and student outcomes indicate a need for rethinking and reframing. For example, greater attention is needed to the following:

I. Variables Relevant to Interpreting Results
- A. Clarification of Mission and Rationale for an Approach
- B. Antecedents & Inputs
- C. Processes & Transactions
- D. Current Outcomes (positive and negative) with reference to trends and goals

II. Content Focus
- A. Cognitive Development and Engagement
- B. Physical and Social Development and Behavioral Engagement
- C. Emotional Development and Engagement
- D. System Performance and Ongoing Development

III. Levels
- A. National
- B. State
- C. School District and Surrounding Community
- D. School and Neighborhood
- E. Classroom
- F. Individuals

Exhibit 3 uses content focus (cognitive, physical, social, behavioral, and emotional development and engagement) as categories for outlining a range of short-term outcome indicators. Examples are offered for each category. In reviewing items, remember that dimensions range from negative to positive.

Again we stress that data related to complex phenomena must be interpreted cautiously and with concern for bias. Remember: a student and school are complex entities that are divided and categorized into multiple theoretical dimensions; categories are constructs; categories overlap; indicator measures have limited construct validity; chosen indicators and available measures capture only a snapshot of reality; all school data requires careful disaggregation; outcomes alone are insufficient for determining cause and effect.

Exhibit 3

Categories and Examples of Short-term Outcome Indicators

I. ***Indicators of Cognitive Development and Engagement***

A. *Maintenance and general application of knowledge* (e.g., evidence of amount learned and use of the learning at school and elsewhere)

B. *Positive behavioral and emotional engagement in acquiring and applying knowledge* (see examples below)

C. *Cognitive coping* (e.g., strategies used at school and elsewhere to learn and apply knowledge and overcome barriers to knowledge acquisition and use)

II. ***Indicators of Physical and Social Development and Behavioral Engagement***

A. *Physical health* (e.g., age-appropriate body and sensory development, safe behaviors)

B. *Personal and social functioning and coping* as manifested in

1. expressed expectations and valuing (e.g., expectations of outcomes; valuing and interest in learning at school; types of choices made when options are available)

2. *conduct* (e.g., acceptance of personal responsibility; rule compliance-noncompliance; completing assignments; attendance; truancy; tardies; referrals for misbehavior; expulsion; suspension; dropping out)

3. *persistence and problem solving* (e.g., effort, attention-inattention, coping-noncoping; grades)

4. *participation* (e.g., in academic activities; in extracurricular activities; in social situations; on-off task; leader-follower; degree of enthusiasm; degree of involvement; initiating-withholding)

III. ***Indicators of Emotional Development and Engagement***

A. *School-related attitudes* (e.g., about school, teachers, peers, schoolwork, self as learner – including *feelings* of competence, self-determination, and relatedness; psychological reactance; perceptions of belonging and being cared about; perception of fairness; feeling safe-victimized; hope for the future)

B. *Other attitudes that may be affecting engagement at school* (e.g., positive and negative feelings related to neighborhood, family, peer, self as a person – including *feelings* of competence, self-determination, and relatedness; feeling safe-victimized)

C. *Emotional coping* (e.g., strategies used at school and elsewhere to respond to affect)

(cont.)

IV. **Indicators of System Performance and Ongoing Development**

A. *All the above can be used as system performance indicators*

B. *Stakeholder groups (students, families, staff, community) perceptions of school culture and climate*

C. *System development for facilitation of cognitive, physical, social, behavioral, and emotional development and engagement* (e.g., status of instruction and curriculum and schoolwide programs for facilitating learning and development)

D. *System of supports to address barriers to learning & teaching and re-engage disconnected students* (e.g., range of interventions; status of development of interventions into a comprehensive system of student and learning supports pre-k through post secondary; integration into school improvement policy and practice)

E. *Development of a school-family-community collaboration for system building to enhance cognitive, physical, social, behavioral, and emotional development and engagement* (e.g., status of policy, operational infrastructure, and capacity building supports for collaboration)

F. *Overall system governance and management* (e.g., status of policy, infrastructure, monitoring and capacity building supports -- including professional and other stakeholder development, cost effectiveness and efficiency)

Why Policy Makers Must Change School Accountability Practices

Accountability indicators have extraordinary power to reshape schools. Systems are driven by what is measured for purposes of accountability. This is particularly so when systems are involved in major reform and transformation.

Under reform conditions, policy makers often want a quick and easy recipe to use. This leads to measures aimed at holding administrators and staff accountable for specific, short-term results. Little thought is given to the negative effects such a limited focus can have on achieving more complex desired long-term results.

A Growing Disconnect

Current school accountability is a good example of the problem. The situation is one where accountability demands focus on a narrow set of outcome indicators. School personnel are quick to learn what will and will not be evaluated, and slowly but surely greater emphasis is placed on teaching what will be measured. Over time what is measured increasingly becomes viewed as the most important outcomes to be achieved (e.g., reading, math, science), and other educational opportunities and essential

student and learning supports are deemphasized and even dropped.

What's wrong with that? Nothing – if what is being evaluated reflects all the important things we want youngsters to learn in school and focuses on enabling equity of opportunity for success at school. This, of course, is not the case.

Prevailing accountability pressures reflect values and biases that are reshaping the entire nature and scope of schooling. As everyone involved in school improvement knows, the only measures that really count are achievement test scores. These scores drive school accountability. What the tests measure has become the be-all and end-all of school improvement policy and planning. This produces a growing disconnect between the direction in which many policy makers and school reformers are leading the public and the realities of what it takes to improve academic performance and student well-being.

The disconnect is especially evident in schools enrolling students from "low wealth" families. Such families and those who work in schools serving them have a clear appreciation of many barriers that must be addressed so students can benefit from the teacher's efforts to teach. These stakeholders stress that major academic improvements are unlikely until comprehensive and multifaceted approaches for addressing the barriers are developed and pursued effectively.

At the same time, anyone who looks will find no direct accountability for addressing barriers to learning and teaching and re-engaging disconnected students. Ironically, the lack of an accountability focus on these matters contributes to devaluing of and justifying cuts in student and learning supports.

Thus, rather than building the type of system that can produce improved academic performance, prevailing accountability measures pressure schools to pursue mainly a direct and ineffective route to improving instruction. The implicit underlying assumption of the direct route is that students are motivationally ready and able each day to benefit from the teacher's instruction. The reality, of course, is that in many schools the majority of youngsters do not fit this picture and are not benefitting from promising instructional improvements. The results of persevering in this direction are continuing low test scores and an ongoing achievement gap.

Logically, major systemic efforts should address interfering factors. However, current accountability pressures override the

logic and result in marginalizing almost every initiative not viewed as a direct and quick path to higher achievement test scores. The irony is that such policy not only works against what must be done, it works against gathering evidence on the necessity and effectiveness of directly and comprehensively addressing barriers to learning.

Needed: An Expanded Accountability Framework

In moving forward, an expanded framework for school accountability is needed. To this end, our Center has emphasized the need for a framework that

- encompasses a whole person approach to student outcomes (i.e., cognitive development and engagement, physical and social development and behavioral engagement, and emotional development and engagement)
- addresses a fuller range of barriers to learning and teaching
- assesses the school's role in strengthening families and neighborhoods
- evaluates system performance and development and does so in the context of the surrounding neighborhood (e.g., Adelman & Taylor, 2006; Center for Mental Health in Schools, 1998).

As to indicators, this brief has highlighted categories and examples relevant to such an expanded accountability framework (see Exhibit 3) and again underscores the matter below.

We view the expanded framework as a move toward what has been called intelligent accountability. The intent is not to deflect from the laser-like focus on accountability for meeting high standards related to academics. The debate will continue as to how best to measure academic outcomes, but clearly schools must demonstrate they effectively teach academics.

Schools also are expected, however, to pursue high standards in *promoting positive social and personal functioning*, including enhancing engagement, civility, teaching safe and healthy behavior, and some form of "character education." Every school we visit has specific goals related to this facet of student development and learning. But, schools currently are not held accountable for goals in this arena. That is, no systematic evaluation or reporting of the work is done. As would be expected, then, schools direct few resources and too little

attention to these unmeasured concerns. Yet, society wants schools to attend to these matters, and most professionals understand that personal and social functioning is integrally tied to academic performance. From this perspective, not holding schools accountable for improving students' social and personal functioning is self-defeating.

For schools where a large proportion of students are not doing well, not attending to benchmark indicators of progress in addressing barriers to learning also is self-defeating. Schools cannot teach children who are not in class. Therefore, increasing attendance always is an expectation (and an important budget consideration). Other basic indicators of school improvement and precursors of enhanced academic performance are reducing tardiness and problem behaviors, lessening suspension and dropout rates, and abating the large number of inappropriate referrals for special education. Given this, the progress of school staff in addressing such problems should be measured and treated as a significant aspect of school accountability.

School outcomes, of course, are influenced by the well-being of the families and the neighborhoods in which they operate. Therefore, performance of any school should be judged within the context of the current status of indicators of community well-being, such as economic, social, and health measures. When those indicators are not improving or are declining, schools find it difficult to make progress. Judging school performance out of context is patently unfair.

Thus, an expanded accountability framework is needed to encourage and support movement toward a broad band approach to addressing barriers to learning and teaching and re-engaging disconnected students. Such a broad approach recognizes the interconnectedness of neighborhood, family, school, and student factors, therefore, changes in all are a relevant focus of data gathering. We are reminded of Ulric Neisser's dictum:

> *Changing the individual while leaving the world alone is a dubious proposition.*

Protocol for an Expanded Accountability Framework

School accountability is a policy tool with extraordinary power to reshape schools – for good and for bad. Systems are driven by accountability measures. This is particularly so under "reform" conditions.

Accountability is a policy tool that has extraordinary power to reshape schools

As everyone involved in school reform knows, the only measure that really counts is achievement test scores. These tests drive school accountability, and what such tests measure has become the be-all and end-all of what is attended to by many decision makers. This produces a growing disconnect between the realities of what it takes to improve academic performance and the direction in which many policy makers and school reformers are leading the public.

The disconnect is especially evident in schools serving what are now being referred to as "low wealth" families. Such families and those who work in schools serving them have a clear appreciation of many barriers to learning that must be addressed so students can benefit from the teacher's efforts to teach. These stakeholders stress that, in many schools, major academic improvements are unlikely until comprehensive and multifaceted approaches to address these barriers are developed and pursued effectively.

At the same time, it is evident to anyone who looks that there is no direct accountability for whether these barriers are addressed. To the contrary, efforts essential for addressing barriers to development and learning are further devalued and cut when achievement test scores do not reflect an immediate impact.

Current accountability pressures marginalize almost every effort not seen as directly and quickly producing higher achievement scores

Thus, rather than building the type of system that can produce improved academic performance, prevailing accountability measures are pressuring schools to pursue a direct route to improving instruction. The implicit underlying assumption is that students are motivationally ready and able each day to benefit from the teacher's instruction. The reality, of course, is that the *majority* of youngsters do not fit this picture in too many schools. Students confronted with a host of external interfering factors usually are not in a position to benefit even from significant instructional improvements. The result is low test scores and an achievement gap.

Logically, well designed, systematic efforts should be directed at addressing interfering factors. However, current accountability pressures override the logic and marginalize almost every effort not seen as directly and quickly leading to higher achievement scores. Ironically, this works against what must be done and against gathering evidence on how the impact of addressing barriers to learning directly.

All this leads to an appreciation of the need for an expanded framework for school accountability – a framework that includes direct measures of achievement and much more. We view this as a move toward what has been called *intelligent accountability.* The following Exhibit highlights such an expanded framework.

As illustrated, there is no intent to deflect from the laser-like focus on meeting high academic standards. Debate will continue about how best to measure academic outcomes, but clearly schools must demonstrate they effectively teach academics.

At the same time, policy must acknowledge that schools also are expected to pursue high standards in promoting positive social and personal functioning, including enhancing civility, teaching safe and healthy behavior, and some form of "character education." Every school we visit has specific goals related to this facet of student development and learning. Yet, it is evident that there is no systematic evaluation or reporting of the work. As would be expected, then, schools direct few resources and too little attention to these unmeasured concerns. Yet, society wants schools to attend to these matters, and most professionals understand that personal and social functioning are integrally tied to academic performance. From this perspective, it seem self-defeating not to hold schools accountable for improving students' social and personal functioning.

It is self-defeating not to attend to benchmark indicators of progress related to addressing barriers to learning

For schools where a large proportion of students are not doing well, it is also self-defeating not to attend to benchmark indicators of progress in addressing barriers to learning. Schools cannot teach children who are not in class. Therefore, increasing attendance always is an expectation (and an important budget consideration). Other basic indicators of school improvement and precursors of enhanced academic performance are reducing tardiness and problem behaviors, lessening suspension and dropout rates, and abating the large number of inappropriate referrals for special education. Given this, the progress of school staff related to such matters should be measured and treated as a significant aspect of school accountability.

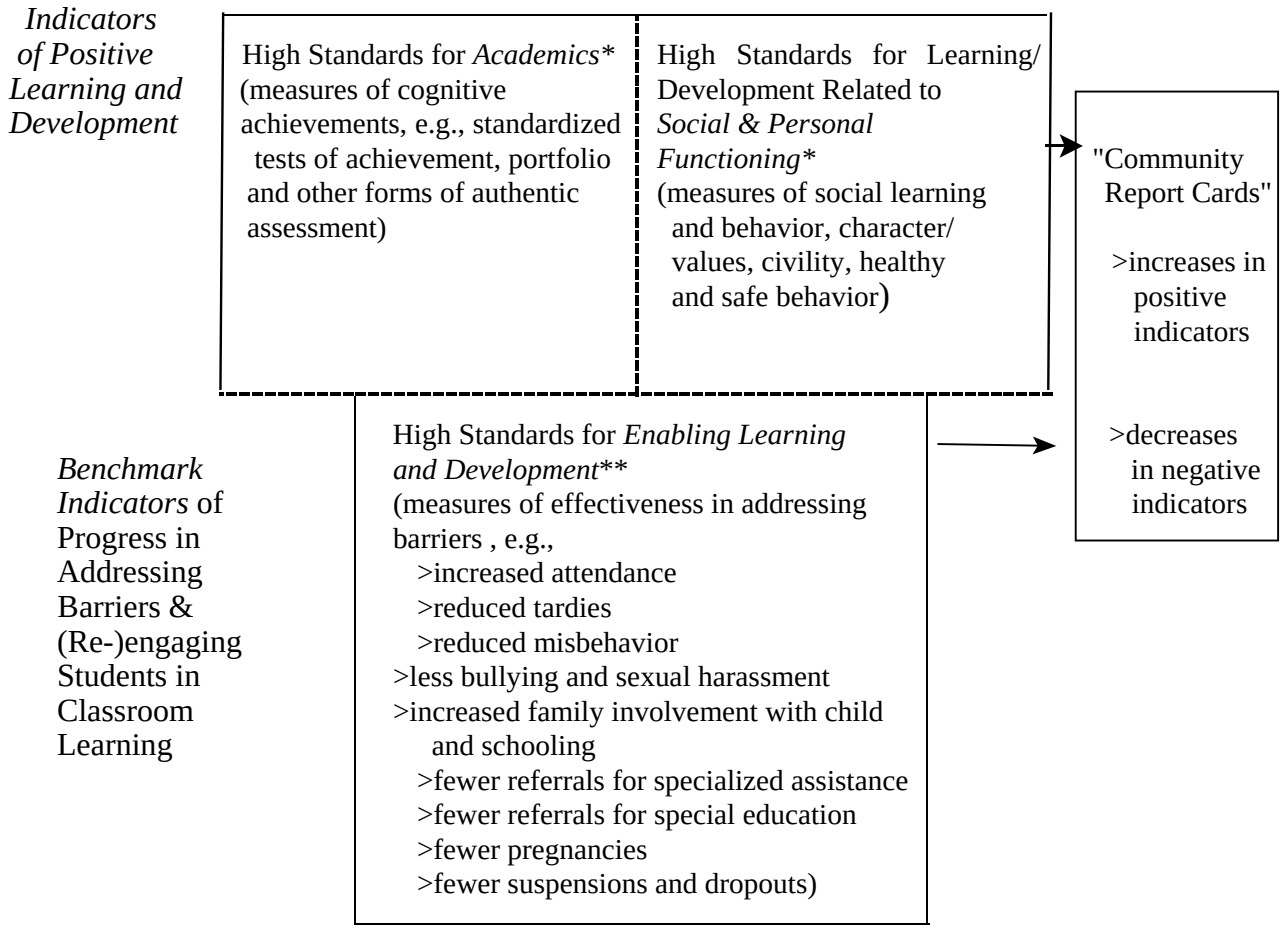
School outcomes, of course, are influenced by the well-being of the families and the neighborhoods in which they operate. Therefore, performance of any school should be judged within the context of

the current status of indicators of community well-being, such as economic, social, and health measures. If those indicators are not improving or are declining, it is patently unfair to ignore these contextual conditions in judging school performance.

In sum, it is unlikely the majority of students in economically depressed areas will perform up to high standards if schools and communities do not pursue a holistic, systemic, and collaborative approach that focuses not just on students, but on strengthening their families, schools, and surrounding neighborhood.

Exhibit

Expanding the Framework for School Accountability



*Results of interventions for directly facilitating development and learning.

**Results of interventions for addressing barriers to learning and development.

A Few Concluding Introductory Comments

> *Today's enthusiastic embrace of data has waltzed us directly from a petulant resistance to performance measures to a reflexive and unsophisticated reliance on a few simple metrics.... The result has been a nifty pirouette from one troubling mind-set to another; with nary a mistep, we have pivoted from the "old stupid" to the "new stupid."*
>
> Frederick Hess

The need for professionals to improve their practices and be accountable is obvious. Gathering good data to evaluate schools clearly contributes to school improvement. Doing so, however, is a more complex problem than just focusing on gathering data related to currently emphasized indicators of results .

Because evaluations can as easily reshape programs in negative as in positive directions such practices must be improved, and accountability pressures must not inappropriately narrow a program's focus. This is especially the case for programs designed to enable the learning of students who are not doing well at school, including new initiatives and specially funded projects such as those related to addressing psychosocial and mental and physical health concerns.

In moving forward, policy makers must do more than mandate a narrow band of accountability. They must

- expand the framework for school accountability to ensure that systems are driven in ways that provide an equal opportunity for all students to succeed at school

- invest in the development of a set of broad band indicators, including a focus on (a) a whole person approach to student outcomes, (b) a fuller range of barriers to learning and teaching, (c) the school's role in strengthening families and neighborhoods, and (d) system performance and development in the context of the surrounding neighborhood

- invest in supporting districts and school development of information management systems that enable gathering and sharing data in aggregated and disaggregated ways (including safeguarded data on individuals).

Cited References

Adelman, H.S., & Taylor, L. (2006). *The school leader's guide to student learning supports: New directions for addressing barriers to learning*. Thousand Oaks, CA: Corwin Press.

Center for Mental Health in Schools (1998). *Accountability: Is it becoming a mantra?* Los Angeles, CA: Author. http://smhp.psych.ucla.edu/account.htm

Horsch, K. (1997). *Indicators: Definition and use in a results-based accountability system*. Cambridge, MA: Harvard Family Research Project. http://www.hfrp.org/publications-resources/browse-our-publications/indicators-definition-and-use-in-a-results-based-accountability-system

Child Trends (2009). *Results and indicators for children: An analysis to inform discussions about promise neighborhoods***.** Washington, D.C.: Author. http://www.childtrends.org/Files//Child_Trends-2009_11_06_FR_PromiseNeigh.pdf

Iowa Department of Education with the Iowa Collaboration for Youth Development (2004). *Enhancing Iowa's systems of supports for development and learning*. Des Moines: Author. http://www.iowa.gov/educate/index.php?option=com_docman&task=doc_download&gid=2588

Isakson, E.A., Higgins, L., Davidson, L., & Cooper, J.L. (2009). *Indicators for social-emotional development in early childhood : A guide for local stakeholders.* New York: National Center for Children in Poverty. http://www.nccp.org/publications/pub_901.html

Isakson, E.A., Higgins, L., Davidson, L., & Cooper, J.L. (2011). *State-level indicators for social-emotional development: Building better systems.* New York: National Center for Children in Poverty. http://www.nccp.org/publications/pub_997.html

National Center for School Engagement (2006a). *Quantifying school engagement: Research report.* Denver: Author. http://www.peecworks.org/PEEC/PEEC_Inst/035FA6B1-007EA7AB.0/Finlay%202006%20Quantifying%20School%20Engagement.pdf

National Center for School Engagement (2006b). *Developing a report card for Triple A schools.* Denver: Author. http://www.schoolengagement.org/TruancypreventionRegistry/Admin/Resources/Resources/DevelopingaReportCardforTripleASchools.pdf

National School Climate Council (2010). *National school climate standards: Benchmarks to promote effective teaching, learning, and comprehensive school improvement.* New York: National School Climate Center. http://www.schoolclimate.org/climate/policy.php

Planty, M. & Carlson, D. (2010). *Understanding education indicators: A practical primer for research and policy.* New York: Teachers College Press.

Ravitz, D. (2010). *The death and life of the great american school system: How testing and choice are undermining education.* New York: Basic Books.

Rutkowski, D. (2008). Towards an Understanding of Educational Indicators. Policy Futures in Education, 6, 470-481. http://www.wwwords.co.uk/pdf/validate.asp?j=pfie&vol=6&issue=4&year=2008&article=9_Rutkowski_PFIE_6_4_web

Shavelson, Richard J., McDonnell, L. & J. Oakes (1991). What are educational indicators and indicator systems? *Practical Assessment, Research & Evaluation, 2.* http://PAREonline.net/getvn.asp?v=2&n=11

Section I C:

Q & A: *What's the Science-Base for What You Do?*

Related to demands for accountability are increasing demands to demonstrate there is a science/ research/evidence-base to support practices conducted in schools. As school budgets are cut, policy makers increasingly ask: *Where's the evidence that the type of work you are doing helps the school meet its educational mission?* And, the inability to provide an answer to this million dollar question often is the precursor to eliminating programs and laying off staff.

The question is so logical and simple to ask. The problem is that such questions imply that relevant data are easy to gather, and so if data aren't available, the intervention must be ineffective. Usually ignored by the questioners are the many complexities associated with valid and ethical evaluation of major efforts to address problems, reengage learners, etc.

As any researcher will acknowledge, fundamental dilemmas stem from the limited validity and focus of available measures and also from the tendency for those demanding evidence to have inappropriate expectations. Most widely sanctioned evaluation instruments are quite fallible. Moreover, they are designed to measure results that require a lengthy course of intervention, thereby giving short shrift to immediate benefits (benchmarks) that are essential precursors of longer-range improvements. This is particularly a concern when problems being addressed are severe, pervasive, and chronic since rapid improvement is improbable.

Given policy- and budget-driven demands for empirical support, it is not surprising that a frequent request our Center receives is for "data or sources of data that indicate the impact of student support staff and learning support interventions on achievement, graduation rates or completion rates or longer stays in school, reengagement in classroom learning of students who have become disengaged, increased SAT scores; increased number of students taking AP classes, and so forth."

Here is how we respond to such a request:

There are many specific reports and articles that can be cited. For example:

>>A quick aid in responding to policy makers is to draw on the recent report from West Ed based on data from the California Healthy Kids Survey. Aspects of it are being used to show a link between school academic performance/progress and student health and resilience – the report is online at http://www.wested.org/cs/we/view/rs_press/22

>>With specific respect to Student Engagement, the research stresses that *Engagement is related positively to achievement; disengagement is a precursor to dropping out of school.* See: "School Engagement: Potential of the Concept, State of the Evidence" (2004) by J. Fredricks, P. Blumenfeld, & A. Paris. *Review of Educational Research, 74*, 59-109.

However, we find that citing specific references seldom is persuasive to policy makers. Thus, we think it is a fundamental mistake for those arguing for the importance of student support staff and learning support interventions to approach the matter in too narrow a manner. Instead we offer a response that reflects our extensive review of the literature encapsulating outcome info from about 200 programs grouped using a comprehensive framework that represents the full range of activity such schools and student support staff provide as learning supports. That "enabling" or "learning support" framework consists of

six basic areas that address barriers to learning and enhance healthy development: (1) enhancing classroom-based efforts to enable learning, (2) providing prescribed student and family assistance, (3) responding to and preventing crises, (4) supporting transitions, (5) increasing home involvement in schooling, and (6) outreaching for greater community involvement and support (including use of volunteers).

The review is provided in an extensive document entitled: "A Sampling of Outcome Findings from Interventions Relevant to Addressing Barriers to Learning" which is online at: http://smhp.psych.ucla.edu/pdfdocs/Sampler/Outcome/outcome.pdf

Because that document is so large, we have developed a very brief document for policy makers that summarizes the research base (entitled: "Addressing Barriers to Student Learning & Promoting Healthy Development: A Usable Research-Base.") This brief has been used in places around the country to satisfy policy makers that the need for learning supports and the staff who provide them is essential to leaving no child behind. Downloadable at: http://smhp.psych.ucla.edu/pdfdocs/briefs/BarriersBrief.pdf

The brief document begins with the following statement which highlights our approach to this matter at this time:

> "As schools evolve their improvement plans in keeping with higher standards and expectations and increased account-ability, most planners recognize they must include a comprehensive focus on addressing barriers to student learning and promoting healthy development.[1-15] This awareness finds support in an extensive body of literature. It is illustrated by a growing volume of research on the value of schools, families, and communities working together to provide supportive programs and services that enable students to learn and teachers to teach.[16-22] Findings include improved school attendance, fewer behavior problems, improved inter-personal skills, enhanced achievement, and increased bonding at school and at home.[23]
>
> Given the promising findings, state and local education agencies all over the country are delineating ways to enhance social, emotional, and behavioral performance as an essential facet of improving academic performance. Among the many initiatives underway are those designed to enhance systems of learning supports to better address barriers to learning and promote healthy development. These initiatives are building on a body of research that clarifies the importance of and bases for comprehensive approaches. This brief highlights the research base for key elements of a comprehensive approach."

Section I D:

Evaluating Mental Health in Schools

>> Quality Control, Evaluating Outcomes, and Getting Credit for All You Do

>> Q & A: Evaluating the Impact on Students of Mental Health in Schools

Mental Health in Schools: Quality Control, Evaluation of Outcomes, and Getting Credit for All You Do

a brief discussion with examples of evaluation indicators

We approach mental health activity in schools as one facet of a comprehensive, integrated approach to addressing barriers to learning and enhancing healthy development. The intent of all such activity, of course, is to enhance outcomes for children and adolescents. However, enhancing outcomes for the large number of those in need of help usually involves addressing the systems that determine such outcomes (e.g., families, education support programs, school-based health centers, off-site services, the community at large). Moreover, it is important to proceed with a holistic perspective (e.g., viewing children in the context of families and communities). Such a perspective fosters appreciation of relationships among individuals, specific aspects of systems, and the system as a whole. Given this comprehensive orientation to mental health, it is evident that evaluation involves more than measuring outcomes for individuals served.

Broadly stated, evaluation should be planned and implemented in ways that measure outcomes and much more with a view to enhancing the quality of intervention efforts and the long-term benefits for students and society. The following sections highlight a few ideas along these lines.

Evaluation that Fosters Quality Improvement

One purpose of outcome evaluation is to provide feedback on efficacy so processes can be revised and fine-tuned. Such *formative* evaluation also includes information on participants, approaches, resources, implementation strategies, program organization, staffing, operational policies and practices. It also should include data on the characteristics of the system's "clients" -- who they are, what they want and need, how they differ from those in other locales -- as a prerequisite for effective planning and as another basis for interpreting the appropriateness of observed processes and outcomes. (That is, it is essential to understand the status of clients before an intervention is implemented, not only to be aware of their needs but ultimately to make appropriate judgments about intervention outcome efficacy.)

Thus, formative evaluation includes data gathering and analyses focused on such matters as

- needs and assets, goals and desired outcomes, resources, and activities
- challenges and barriers to mental health intervention and the integration of such interventions with other activity designed to address barriers to learning, as well as with the instructional and management components of schools and communities
- characteristics of families and children in each locale, with special focus on targeted groups
- initial outcomes.

Formative evaluation data may be gathered on and from samples of all parties who have a stake in the intervention (e.g., school staff, students and their families, other stakeholders, community agencies, and so forth). The information is used to judge the "fit" of prerequisite conditions and processes. Methods used include review of documents and records, checklists, surveys, semi-structured interviews, focus group discussions, observations, and direct assessment of clientele. A well-designed information management system can be a major aid (e.g., providing data on identified needs and current status of individuals and resources). In this respect, an advanced technology can play a major role (e.g., a computerized system that is properly designed can provide access to information in other computer-based data systems containing relevant information on clients and processes).

To be maximally useful, a data set should allow for baseline and subgroup comparisons and include multiple variables so that findings can be desegregated during analysis Of particular interest are

data differentiating clients in terms of demographics, initial levels of motivation and development, and type, severity, and pervasiveness of problems. With respect to process, it is useful to have data differentiating stages of program development and differences in program quality.

Optimally, the date gathered should allow for formative-leading-to-summative evaluations. Designing s formative evaluation system that over time yields summative findings facilitates ongoing planning in ways that improve processes and thus outcomes. At the same time, such an approach builds a system for validating interventions.

Evaluation Focused on Results

To begin with, it will help to clarify our definition of some terms that are used throughout this section. *Aims* are extremely abstract statements of intended outcomes that encompass many goals and objectives; this usually means an aim can only be accomplished over an extensive time period (e.g., many years). *Goals* are somewhat less abstract statements encompassing many objectives; thus, a goal usually requires a somewhat extended period of time to accomplish. *Objectives* are meant to be less abstract and more immediately accomplishable than the goal that encompasses them. A *standard* is defined as a statement about what is valued. Standards are used to (a) judge and promote quality, (b) clarify goals, and (c) promote change. In evaluating efficacy, standards are operationalized in terms of specific *criteria* upon which judgments of immediate and potential long-term efficacy can be made. *Indicators of efficacy* are measurable variables that can be accessed from various sources through use of specific data gathering strategies and tools.

As emphasized above, while the intent of mental health activity in schools is to enhance outcomes for students, the effort must also address the systems that determine such outcomes. Thus, the following discussion outlines intended impact not only on students, but on families and community, and on programs and systems.

Student Outcomes

Efforts to address mental health concerns and other barriers to learning include enhancing receptivity to instruction through facilitating positive academic, social, emotional, and physical development. In this section, we focus first on outcomes related to facilitating such development; then, the emphasis shifts to prevention and correction of emotional, behavioral, learning, and health problems.

(1) *Outcomes reflecting enhanced receptivity to instruction.* Teaching and learning are transactional. Students (and teachers) bring certain capacities and attitudes (abilities, expectations, values) accumulated and established over time. These provide the foundation upon which teaching tries to build. Students also come with current physiological and psychological states of being that can facilitate or inhibit learning at any given time. Efforts to enhance receptivity to instruction focus on ensuring there is a good instructional match with the student's capacities, attitudes and current state of being. While this is especially necessary for those manifesting serious problems, it is a fundamental concern related to all learners.

The *aim* of enhancing receptivity to instruction involves ensuring that students have the opportunity to acquire the types of basic abilities, expectations, and values that enable learning. The aim also encompasses the need for schools to respond appropriately to variations in students' current states of being (e.g., ensuring the opportunity to learn by providing breakfast and lunch programs to combat hunger, responding to personal problems and crises with support and guidance).

As is highlighted by the goals and objectives outlined in Exhibit A, the ultimate aim is to ensure that students develop effective levels of functionality -- academically, socially, emotionally, and physically. (With respect to social-emotional functioning, aims are sometimes referred to as personal qualities, interpersonal functioning, the affective domain, and so forth. Physical functioning often is discussed as physical and health education.) From a developmental perspective, the aim encompasses concerns for ensuring a "healthy start," a safe school environment, preparation

(readiness) for school, facilitating continued positive development in all areas, facilitating progress with respect to developmental tasks at each stage of development, enhancing areas of personal interest and strength, and fostering a psychological sense of community. As with all curricular goals, desired outcomes in these areas reflect (a) intended uses (communication, reasoning, problem solving, making relationships and connections, and creativity) and (b) factors related to intrinsic motivation (personal valuing and expectations of efficacy -- including confidence in one's abilities).

The goals and objectives outlined in Exhibit A provide a frame of reference for designing programmatic activity to facilitate development related to enhancing receptivity to instruction through facilitating positive academic, social, emotional, and physical development. It is clear that attending to such functioning is basic to preventing, treating, and remedying problems. Moreover, the goals and objectives provide direction for daily program planning and for evaluation.

The assumption in pursuing goals and objectives is that optimal processes (comprehensive and integrated programs) will be used to create a match that enhances positive attitudes, growth, and learning. This applies to the full range of support available to students and families -- including specialized programs at the site, home, and community. Until a comprehensive, integrated continuum of programs and services are in place, steps must be taken to address the less than optimal conditions. From this perspective, evaluation focuses on (a) individual student outcomes (related to the goals and objectives set forth in Exhibit A) and (b) outcomes for all children in the catchment area (e.g., community indicators of improved health, safety and survival, emotional health, and positive social connections). In addition, there can be a focus on outcomes reflecting significant changes in support systems (e.g., measures of enhanced home involvement in schooling; indicators of enhanced integration of center and community health, social, and mental health services – including related data on financial savings).

Furthermore, in pursuing goals and objectives related to instructional receptivity and social-emotional and physical development, it is essential to do so in ways that value and foster rather than devalue and inhibit appropriate diversity among students. This is especially important given the diversity students bring with regard to ethnic background, gender, interests, and capabilities. Thus, another focus for evaluation is on these concerns (especially in assessing for negative outcomes). In particular, efforts should be made to measure (a) movement toward inappropriate conformity in thinking and behaving in areas where diversity is desired and (b) trends toward increased levels of other-directedness and excessive dependency.

(2) ***Outcomes related to preventing and correcting emotional, behavioral, learning, and health problems.*** In addition to the above goals and objectives, student goals and objectives are formulated in connection with specialized programs designed to prevent and correct emotional, behavioral, learning, and health problems. These objectives relate to the efforts of such programs to remove barriers and enable students to pursue the above goals.

It is important to emphasize that problems become of concern because they are reflected in the student's functioning; however, the primary source of the problem often is environmental. Environmentally based problems are an especially important focus for prevention programs. Such programs are targeted to designated at-risk populations (e.g., students with older siblings in gangs, immigrant and highly mobile families who have major transition and school adjustment needs, students who experience a crisis event).

In general, then, immediate objectives in working to address emotional and behavioral problems with a view to enabling student progress often include activity designed to reduce specified barriers to school attendance and functioning. Thus, attending to mental health concerns often requires addressing practical deterrents such as health problems, lack of adequate clothing, problems in the home, working with home to increase support for student improvement, dealing with student's physical or sexual abuse, dealing with student's substance abuse, dealing with gang involvement, provisions for pregnant minors and minor parents, dropout outreach and recovery, teaching student to use compensatory strategies for learning, and so forth. And, based on the discussion to this point, hopefully it is clear that the first indicators of progress may be fewer problems related to learning,

behavior, and affect. See Exhibit A for examples of key intervention goals and objectives and potential indicators of efficacy. The goals and objectives listed in Exhibit A represent individual student outcomes that can be measured as indicators of the impact of specialized programs. Positive "side effect" outcomes worth measuring are significant changes related to (a) all children in the catchment area (e.g., community indicators of improved health, safety and survival, emotional health, and positive social connections) and (b) support systems (e.g., enhanced home involvement in schooling; enhanced integration of a school-based health center and community health, social, and mental health services -- including related data on financial savings). Of course, additional student outcomes can be delineated and measured with respect to efforts to prevent specific types of problems. This is usually accomplished by fostering positive functioning through activities designed to enhance knowledge, skills, attitudes, and action related to healthy physical and mental development. Some of these efforts are carried out in special settings, such as school-based health centers and family resource centers. Whether or not there is a special setting, these efforts include specialized programs focused on

- home involvement to enhance social-emotional development
- peer-to-peer interventions designed to enhance social-emotional development
- early education for prenatally drug-exposed children and their families
- substance abuse prevention
- suicide prevention
- physical and sexual abuse prevention
- violence prevention
- dropout prevention and school re-entry
- STD/AIDS prevention
- pregnancy prevention
- prenatal care of pregnant minors and minor parent education
- crisis intervention and emergency responses to prevent long-term impact (e.g., PTSD) and to prevent subsequent emergencies

Intended Impact on Families and Community

Aims related to families encompass promotion of positive family development and functioning and enhanced home involvement in schooling. Aims for the community encompass promotion of positive community development and functioning and related reform of community agencies (with particular emphasis on reducing problems related to health and safety). See Exhibit B for examples of key intervention goals and objectives and potential indicators of efficacy.

Intended Impact on Programs and Systems

Major aims with respect to the school-site are to promote and support (a) a major restructuring of school support services, (b) integration of school support services with other school-based/linked support programs, teams, and special projects (in both the regular and special education arenas), (c) outreach to enhance linkages and collaborations with community resources (e.g., health, social, recreational programs; involvement of volunteers and local businesses), and (d) integration of all activity designed to address barriers to learning with the instructional and school management components. See Exhibit C for examples of key goals and objectives and of potential indicators of efficacy.

Exhibit A

Intervention Impact on Students

Aims	Examples of Goals/Objectives	Examples of Indicators of Efficacy	Standards/Criteria Immediate -- Long-term
Enhance receptivity to instruction Prevent and correct emotional, behavior, learning, & health problems	Increase knowledge, skills, & attitudes to enhance •acceptance of responsibility (including attending, following directions & agreed upon rules/laws) •self-esteem & integrity •social & working relationships •self-evaluation & self-direction/regulation •physical functioning •health maintenance •safe behavior Reduce barriers to school attendance and functioning by addressing problems related to •health •lack of adequate clothing •dysfunctional families •lack of home support for student improvement •physical/sexual abuse •substance abuse •gang involvement •pregnant/parenting minors •dropouts •need for compensatory learning strategies	Ratings by staff, family, peers Self-reports by students Performance indices (focus is on: •readiness/prerequisites/ survival skills •attendance •tardies •distractibility/ daydreaming/overactivity •dependence on others in pursuing tasks and controlling behavior •misbehavior •symptoms •negative attitudes toward self, teachers, school, peers, family, society) (Ultimately, of course, a major focus is on grades and achievement test scores.)	TO BE DETERMINED BY SITE In developing standards and criteria, it may be helpful to review the report from the Policy Leadership Cadre for Mental Health in Schools – *Mental Health in Schools: Guidelines, Models, Resources, and Policy Considerations* online at http://smhp.psych.ucla.edu/pdfdocs/policymakers/cadreguidelines.pdf.

Exhibit B

Intervention Impact on Families and Communities

Aims	Examples of Goals/Objectives	Examples of Indicators of Efficacy	Standards/Criteria Immediate -- Long-term
Promotion of positive family development & functioning Enhanced home involvement in schooling	Increase social and emotional support for families Increase family access to special assistance Increase family ability to reduce child risk factors that can be barriers to learning Increase bilingual ability and literacy of parents Increase family ability to support schooling Increase positive attitudes about schooling Increase home (family/parent) participation at school	Parents rate satisfaction with school & community programs & services designed to enhance family functioning & provide assistance Staff rates functioning of families Frequency counts of services/ programs in operation; Performance indices Staff rates functioning of families Family self-reports Frequency counts of areas of participation and number of participants	TO BE DETERMINED BY SITE In developing standards and criteria, it may be helpful to review the report from the Policy Leadership Cadre for Mental Health in Schools – *Mental Health in Schools: Guidelines, Models, Resources, and Policy Considerations* online at http://smhp.psych.ucla.edu/pdfdocs/policymakers/cadreguidelines.pdf

Exhibit B (cont.)

Intervention Impact on Families and Communities

Aims	Examples of Goals/Objectives	Examples of Indicators of Efficacy	Standards/Criteria Immediate -- Long-term
Promotion of positive community development and functioning (including influencing restructuring of community agencies)	Enhance positive attitudes toward school and community Increase community participation in school activities Increase perception of the school as a hub of community activities Increase partnerships designed to enhance education & service availability in community Enhance coordination & collaboration between community agencies and school programs & services Enhance focus on agency outreach to meet family needs Increase psychological sense of community	Self-reports of community residents Frequency counts of areas of participation and number of participants Self-reports of community residents Existence of partnership agreements & shared decision making mechanisms Staff rates quality of coordination mechanisms & working relationships Frequency counts of students and families using programs and services Self-reports of community residents Data from records on (a) violent acts (b) nonviolent crime (c) public health problems	TO BE DETERMINED BY SITE In developing standards and criteria, it may be helpful to review the report from the Policy Leadership Cadre for Mental Health in Schools – *Mental Health in Schools: Guidelines, Models, Resources, and Policy Considerations* online at http://smhp.psych.ucla.edu/pdfdocs/policymakers/cadreguidelines.pdf

Exhibit C

Intervention Impact on Programs and Systems

Aims	Examples of Goals/Objectives	Examples of Indicators of Efficacy	Standards/Criteria Immediate -- Long-term
Promote and support restructuring of support services (including integration with instruction & management)	Enhance processes by which staff and families learn about available programs and services and how to access those they need	Frequency counts of students and families using programs and services	TO BE DETERMINED BY SITE In developing standards and criteria, it may be helpful to review the report from the Policy Leadership Cadre for Mental Health in Schools – *Mental Health in Schools: Guidelines, Models, Resources, and Policy Considerations* online at http://smhp.psych.ucla.edu/pdfdocs/policymakers/cadreguidelines.pdf
	Increase coordination among services and programs	Staff rates quality of coordination mechanisms	
	Increase the degree to which staff work collaboratively and programmatically	Supervisors and staff rate how staff spends time	
	Increase services/programs at school site	Frequency counts of services/ programs in operation	
Promote and support outreach to community resources & their integration with school programs & services	Increase amount of school and community collaboration	Existence of interagency agreements & shared decision making mechanisms	
	Increase quality of services and programs by improving systems for requesting, accessing, and managing assistance for students and families (including overcoming inappropriate barriers to confidentiality)	Staff rates quality of (a) systems for triage, referral, case monitoring & management; (b) staff development Users rate satisfaction	
	Establish a long-term financial base	Data from financial records	

Q & A: Evaluating the Impact on Students of Mental Health in Schools

Question: *I need help with structured assessments to be used as pre and post measurements that can be used by school staff and/or school mental health professionals. What measurements are being used now in other school mental health programs?*

Response: Since evaluating mental health interventions is difficult and evaluating mental health in schools adds more complexity, our response to this begins with offering some basics to guide gathering broadly-focused impact evaluation data.

1. *Focus*. Minimally, there are two areas of focus in gathering impact data related to mental health in schools – the student and the school.[1]

Regarding the **Student,** the focus is on indicators of

>symptom reduction
>positive development (capabilities and attitudes)
>improved behavior at school (reduced misbehavior and tardiness, increased attendance)
>academic improvement

Regarding the **School,** the focus is on indicators of

>how many are doing better behaviorally and academically and to what degree (with respect to disaggregated subgroups)
>fewer inappropriate referrals for special assistance/special education
>fewer suspensions, expulsions, dropouts
>improved school/classroom climate
>increased family involvement (with child, with schooling)
>cost-effectiveness

2. *Sources of Data*. The most common sources are:

>Student
>Special intervener(s)
>Parent /Family
>Teachers/Staff
>Peers
>School Records

3. *Pre and post measures.* Examples of established instruments are in an accompanying resource to this intro packet, the Center has developed a TA packet entitled: *Evaluation and Accountability Related to Mental Health in Schools.* That resource is organized around frequently asked questions and center responses.

For refferences to specific measures see Section III F.

[1]Descriptors of the student (e.g., demographics, referral information, diagnosis if applicable), characteristics of services provided (type of intervention, number of visits, and provider, fees, payer), and anything about the school that makes it different from others (low performing, urban, rural, etc.) all are needed for various purposes. These include planning, reporting, billing, and accountability, and carrying out such functions with disaggregated data.

II. Sample of Current Resources and Recent References

The evaluation literature is immense. Here are some examples. Each is a gateway to other references and resources.

A Few References

Research and Evaluation in Education and Psychology (4th ed.). D.M. Merten (2015). Thousand Oaks: Sage.

Counseling and Educational Research; Evaluation and Application (3rd ed.). Rick A. Houser (2015). Thousand Oaks: Sage.

Evaluation Theory, Models, and Applications (2nd ed.). D.L. Stufflebeam & C.L.S. Coryn (2014). San Francisco: Jossey-Bass,

Designing Evaluations 2012 Revision Washington, DC: U.S. Government Accountability Office. – http://www.gao.gov/assets/590/588146.pdf

Really New Directions in Evaluation: Young Evaluators' Perspectives: New Directions for Evaluation, Number 131. Sandra Mathison (Ed.). (2011). San Francisco: Jossey-Bass

Resource Sources

Evaluation Resources from the Institute of Museum and Library – http://www.imls.gov/research/evaluation_resources.aspx

University of Michigan – "My Environmental Education Evaluation Resource Assistant" (MEERA) is an online "evaluation consultant" created to assist with user evaluation needs. It points users to helpful resources for evaluating education programs. – http://meera.snre.umich.edu/

University of North Carolina, Greensboro – Program Evaluation Resource Center – http://erm.uncg.edu/oaers/methodology-resources/program-evaluation/

CDC's Framework for Program Evaluation in Public Health has provided a set of steps and standards for practical evaluation by programs and partners. While the focus is public health programs, the approach can be generalized to any evaluation effort. – http://www.cdc.gov/eval/framework/index.htm

CDC's list of other Evaluation Resources – http://www.cdc.gov/eval/resources/index.htm

A basic introduction to evaluation concepts – http://www.socialresearchmethods.net/kb/evaluation.php An online course on evaluation for non-researchers: http://nrepp.samhsa.gov/Courses/ProgramEvaluation/NREPP_0401_0010.html - See more at: http://arts.gov/grants-organizations/art-works/program-evaluation-resources#sthash.F6Rn3TFz.dpuf

Federal Evaluators – http://www.fedeval.net/books.htm

Basic Guide to Program Evaluation (Including Outcomes Evaluation) – http://managementhelp.org/evaluation/program-evaluation-guide.htm

On-Line Evaluation Resource Library (OERL) – http://oerl.sri.com/

Evaluation Handbook W.K. Kellogg Foundation – http://www.wkkf.org/resource-directory/resource/2010/w-k-kellogg-foundation-evaluation-handbook

Planning and Monitoring Evaluation Checklists – http://www.wmich.edu/evaluation/checklists

CIHLC Evaluation Primer: Overview of Relevant Frameworks and Tools – University of British Columbia – http://www.ipe.utoronto.ca/sites/default/files/CIHLC%20Evaluation%20Primer.pdf

Special education evaluation: an overview – http://www.greatschools.org/gk/articles/special-education-evaluation-an-overview/

Also see the Center's Online Clearinghouse Quick Finds on:

>*Evaluation of Programs to Address Barriers to Learning* -- http://smhp.psych.ucla.edu/qf/evaluation.htm

>*Data Management Systems for Schools and Clinics* – http://smhp.psych.ucla.edu/qf/datasystems.htm

>*Cost-Benefit Analyses Relevant to Addressing Barriers to Learning and Mental Health in Schools* – http://smhp.psych.ucla.edu/qf/costbenefitanalysis.htm

>*Empirically Supported Interventions for Children's Mental Health* – http://smhp.psych.ucla.edu/qf/ests.htm

A Few Centers Focusing on Evaluation

The Evaluation Center – Western Michigan University – http://www.wmich.edu/evalctr/

National Center on Educational Outcomes (NCEO) – University of Minnesota – http://education.umn.edu/NCEO/

National Center for Education Statistics – http://nces.ed.gov/help/

The Center for the Study of Testing, Evaluation, and Educational Policy (CSTEEP) – Boston College – http://www.csteep.bc.edu/

The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) –UCLA – http://www.cse.ucla.edu/

National Study of School Evaluation – http://www.nsse.org/

Assessment and Evaluation on the Internet – http://www.ericfacility.net/ericdigests/ed385609.html

BJA Evaluation Website – http://www.bja.evaluationwebsite.org/html/useful_links/index.html

Resources for Methods in Evaluation and Social Research – http://gsociology.icaap.org/methods

Buros Institute of Mental Measurement – http://www.unl.edu/buros/

III. Tools for Evaluating Planning and Implementation

A. The Program Manager's Guide to Evaluation (2nd ed)
(From the Office of Planning, Research and Evaluation (OPRE),
Administration for Children and Families (ACF

B. Two Sample Evaluation Checklists

C. Program Evaluation Standards

D. Assessment

E. Measures Relevant for Accountability to
Specific Youngsters and Families

Section III A:

The Program Manager's Guide to Evaluation, Second Edition

Download: http://www.acf.hhs.gov/sites/default/files/opre/program_managers_guide_to_eval2010.pdf

From: The Office of Planning, Research and Evaluation (OPRE), a unit within the Administration for Children and Families (ACF) which is responsible for advising the Assistant Secretary for Children and Families on increasing the effectiveness and efficiency of programs to improve the economic and social well-being of children and families. OPRE is responsible for performance management for ACF, conducts research and policy analyses, and develops and oversees research and evaluation projects to assess program performance and inform policy and practice. The Office provides guidance, analysis, technical assistance, and oversight to ACF programs on: strategic planning; performance measurement; research and evaluation methods; statistical, policy, and program analysis; and synthesis and dissemination of research and demonstration findings.

As with the original edition of The Program Manager's Guide to Evaluation, this updated edition explains what program evaluation is, why evaluation is important, how to conduct an evaluation and understand the results, how to report evaluation findings, and how to use evaluation results to improve programs that benefit children and families. This second edition has been updated, under the guidance of Kathleen Dwyer, to reflect currently accepted practices, up-to-date terminology, and issues to consider at this time. Tips, samples, and worksheets that were missing from the online version have been inserted back into the guide, as has a thoroughly updated appendix containing a comprehensive list of evaluation resources. Based on feedback within ACF, we have chosen to focus on a single guide that includes examples that would be relevant to all ACF program areas, rather than create separate handbooks for each program.

Readers may also be interested in a related resource produced by the Children's Bureau: *Cost Analysis in Program Evaluation: A Guide for Child Welfare Researchers and Service Providers* (2013)

B. Two Sample Checklists to Aid in the Planning and Implementation Of Evaluations

Sample Evaluation Checklist #1:

Checklist for Program Evaluation Planning

Written by Carter McNamara, PhD

This checklist was obtained from the The Management Assistance Program for Nonprofits, and applies to both nonprofit and for-profit organizations unless noted

Name of Organization: _____

Name of Program: _____

Purpose of Evaluation?

What do you want to be able to decide as a result of the evaluation? For example:

__Understand, verify or increase impact of products or services on customers/clients (e.g., outcomes evaluation)
__Improve delivery mechanisms to be more efficient and less costly (e.g., process evaluation)
__Verify that we're doing what we think we're doing (e.g., process evaluation)
__Clarify program goals, processes and outcomes for management planning
__Public relations
__Program comparisons (e.g., to decide which should be retained)
__Fully examine and describe effective programs for duplication elsewhere
__Other reason(s)_____

Audience(s) for the Evaluation?

Who are the audiences for the information from the evaluation? For example:

__Clients/customers
__Funders/Investors
__Board members
__Management
__Staff/employees
__Other(s)_____

What Kinds of Information Are Needed?

What kinds of information are needed to make the decision you need to make and/or enlighten your intended audiences? For example, information to understand:

__The process of the product or service delivery (its inputs, activities and outputs)
__ The customers/clients who experience the product or service
__Strengths and weaknesses of the product or service
__Benefits to customers/clients (outcomes)
__How the product or service failed and why, etc.
__Other type(s) of information?_____

Type of Evaluation?

Based on the purpose of the evaluation and the kinds of information needed, what types of evaluation is being planned?

__Goal-based
__Process-based
__Outcomes-based
__Other(s)_____

Where Should Information Be Collected From?

__Staff/employees
__Clients/customers
__Program documentation
__Funders/Investors
__Other(s)_____

How Can Information Be Collected in Reasonable and Realistic Fashion?

__questionnaires
__interviews
__documentation
__observing clients/customers
__observing staff/employees
__conducting focus groups among_____
__other(s)_____

When is the Information Needed? _____

What Resources Are Available to Collect the Information? _____

Sample Evaluation Checklist #2:

Program Evaluation: Focusing on Evaluation

This program evaluation checklist was generated by the National Network for Child Care

What are you going to evaluate?

1. What is the purpose of the evaluation?

2. Who will use the evaluation? How will they use it?

Who/users:	How will they use the information?

3. What questions will the evaluation seek to answer? What information do you need to answer the questions?

What I wish to know:	Indicators - How I will know it?

4. When is the evaluation needed?

5. What resources do you need?

• Time available to work on evaluation:

• Money:

• People: professionals, para-professionals, volunteers, participants

Collecting the Information

1. What sources of information will you use?

• Existing information:

• People:

• Observations:

• Pictorial records:

2. What data collection method(s) will you use?
___survey
___interview
___observation
___group techniques
___case study
___tests
___photos/videos
___document review
___testimonials
___expert panel
___simulated problems or situation
___journal, log diary
___unobtrusive measures
___other (list):

3. Instrumentation: What is needed to record the information?

4. What data collection procedures will be used?

5. When will you collect data for each method you've chosen?

Method:	Before Program	During Program	Immediately After	Later

6. Will a sample be used?
___No
___Yes
If yes, describe the procedure you will use:

7. Who will collect the data?

Using the Information

1. How will the data be analyzed?

• Data analysis methods:

• Who is responsible:

2. How will the information be interpreted and by whom?

3. How will the evaluation be communicated and shared?

Managing the Evaluation

1. Implementation plan:

• time line and responsibilities

• management chart

• budget

C. The Program Evaluation Standards

ERIC/AE Digest.

THIS DIGEST WAS CREATED BY ERIC, THE EDUCATIONAL RESOURCES INFORMATION CENTER.

Sound evaluations of educational programs, projects, and materials in a variety of settings should have four basic attributes:

1. Utility
2. Propriety
3. Feasibility
4. Accuracy

"The Program Evaluation Standards," established by sixteen professional associations, identify evaluation principles that when addressed should result in improved program evaluations containing the above four attributes. What follows is a summary of the Standards.

Guidelines and illustrative cases to assist evaluation participants in meeting each of these standards are provided in the full report (Joint Committee, 1994). The illustrative cases are based in a variety of educational settings that include schools, universities, medical and health care fields, the military, business and industry, the government, and law.

UTILITY

The utility standards are intended to ensure that an evaluation will serve the information needs of intended users.

U1 ***Stakeholder Identification***. Persons involved in or affected by the evaluation should be identified, so that their needs can be addressed.

U2 ***Evaluator Credibility***. The persons conducting the evaluation should be both trustworthy and competent to perform the evaluation, so that the evaluation findings achieve maximum credibility and acceptance.

U3 ***Information Scope and Selection***. Information collected should be broadly selected to address pertinent questions about the program and be responsive to the needs and interests of clients and other specified stakeholders.

U4 ***Values Identification***. The perspectives, procedures, and rationale used to interpret the findings should be carefully described, so that the bases for value judgments are clear.

U5 ***Report Clarity***. Evaluation reports should clearly describe the program being evaluated, including its context, and the purposes, procedures, and findings of the evaluation, so that essential information is provided and easily understood.

U6 ***Report Timelines and Dissemination***. Significant interim findings and evaluation reports should be disseminated to intended users, so that they can be used in a timely fashion.

U7 ***Evaluation Impact***. Evaluations should be planned, conducted, and reported in ways that encourage follow-through by stakeholders, so that the likelihood that the evaluation will be used is increased.

FEASIBILITY

The feasibility standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal.

F1 ***Practical Procedures***. The evaluation procedures should be practical, to keep disruption to a minimum while needed information is obtained.

F2 ***Political Viability***. The evaluation should be planned and conducted with anticipation of the different positions of various interest groups, so that their cooperation may be obtained, and so that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted.

F3 ***Cost Effectiveness***. The evaluation should be efficient and produce information of sufficient value, so that the resources expended can be justified.

PROPRIETY

The propriety standards are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results.

P1 ***Service Orientation***. Evaluations should be designed to assist organizations to address and effectively serve the needs of the full range of targeted participants.

P2 ***Formal Agreements***. Obligations of the formal parties to an evaluation (what is to be done, how, by whom, when) should be agreed to in writing, so that these parties are obligated to adhere to all conditions of the agreement or formally to renegotiate it.

P3 ***Rights of Human Subjects***. Evaluations should be designed and conducted to respect and protect the rights and welfare of human subjects.

P4 ***Human Interactions***. Evaluators should respect human dignity and worth in their interactions with other persons associated with an evaluation, so that participants are not threatened or harmed.

P5 ***Complete and Fair Assessment***. The evaluation should be complete and fair in its examination and recording of strengths and weaknesses of the program being evaluated, so that strengths can be built upon and problem areas addressed.

P6 ***Disclosure of Findings***. The formal parties to an evaluation should ensure that the full set of evaluation findings along with pertinent limitations are made accessible to the persons affected by the evaluation, and any others with expressed legal rights to receive the results.

P7 ***Conflict of Interest***. Conflict of interest should be dealt with openly and honestly, so that it does not compromise the evaluation processes and results.

P8 ***Fiscal Responsibility***. The evaluator's allocation and expenditure of resources should reflect sound accountability procedures and otherwise be prudent and ethically responsible, so that expenditures are accounted for and appropriate.

ACCURACY

The accuracy standards are intended to ensure that an evaluation will reveal and convey technically adequate information about the features that determine worth of merit of the program being evaluated.

A1 ***Program Documentation***. The program being evaluated should be described and documented clearly and accurately, so that the program is clearly identified.

A2 ***Context Analysis***. The context in which the program exists should be examined in enough detail, so that its likely influences on the program can be identified.

A3 ***Described Purposes and Procedures***. The purposes and procedures of the evaluation should be monitored and described in enough detail, so that they can be identified and assessed.

A4 ***Defensible Information Sources***. The sources of information used in a program evaluation should be described in enough detail, so that the adequacy of the information can be assessed.

A5 ***Valid Information***. The information gathering procedures should be chosen or developed and then implemented so that they will assure that the interpretation arrived at is valid for the intended use.

A6 ***Reliable Information***. The information gathering procedures should be chosen or developed and then implemented so that they will assure that the information obtained is sufficiently reliable for the intended use.

A7 ***Systematic Information***. The information collected, processed, and reported in an evaluation should be systematically reviewed and any errors found should be corrected.

A8 ***Analysis of Quantitative Information***. Quantitative information in an evaluation should be appropriately and systematically analyzed so that evaluation questions are effectively answered.

A9 ***Analysis of Qualitative Information***. Qualitative information in an evaluation should be appropriately and systematically analyzed so that evaluation questions are effectively answered.

A10 ***Justified Conclusions***. The conclusions reached in an evaluation should be explicitly justified, so that stakeholders can assess them.

A11 ***Impartial Reporting***. Reporting procedures should guard against distortion caused by personal feelings and biases of any party to the evaluation, so that evaluation reports fairly reflect the evaluation findings.

A12 ***Metaevaluation***. The evaluation itself should be formatively and summatively evaluated against these and other pertinent standards, so that its conduct is appropriately guided and, on completion, stakeholders can closely examine its strengths and weaknesses.

Approved by the American National Standards Institute as an American National Standard.
Approval date: March 15, 1994.

FURTHER READING

Hansen, J.B. & Patton, M.Q.(1994). The Joint Committee on Standards for Educational Evaluation's "The Program Evaluation Standards: How to Assess Evaluations of Educational Programs" Book Review. "Educational and Psychological Measurement," 54(2), 550-67.

Joint Committee on Standards for Educational Evaluations, The (1994). "The Program Evaluation Standards: How to Assess Evaluations of Educational Programs." Thousand Oaks, CA: Sage Publications (available from ERIC/AE).

Stufflebeam, Daniel L.(1987) Professional Standards for Assuring the Quality of Educational Program and Personnel Evaluations. "International Journal of Educational Research," 11(1), 125-43.

Thompson, B. (1993) The Revised Program Evaluation Standards and Their Correlation with the Evaluation Use Literature. Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1993) ED 370 999.

ERIC Identifier: ED385612
Publication Date: 1995-10-00
Source: ERIC Clearinghouse on Assessment and Evaluation Washington DC.

D. Assessment

There is a diverse range of assessment modalities commonly utilized for evaluation research. This document, an excerpt from the following reference, summarizes many of these modalities: Wilde, J. & Sockey, S. (1995). *The Evaluation Handbook* (pp. 17-20). Albuquerque, NM: Evaluation Assistance Center - Western Region.

Assessment systems are key to a good evaluation. The overall purpose of an assessment system is to initiate and maintain discussion about how the program addresses the needs of all participants. As part of this, the program staff must be prepared to assess their own effectiveness as well as participant needs and outcomes. In general, an assessment system should lead directly to the evaluation by ensuring measurement at three times throughout the program:

a) A needs assessment will determine the current status of participants' (and potential participants') expertise and knowledge. A needs assessment allows program planners to determine the needs, desires, and goals of the potential participants and/or their parents, teachers, and other stakeholders. The basic questions are, "Where are we now? What do we know about what these students need, what areas are lacking, and what should we address first?"

b) On-going measures of progress will determine the successful features of the program, the shortcomings of the program, and whether program implementation and the participants are progressing in the expected manner. Measures of progress allow staff to determine whether the program is working and allow participants to see their own growth. The basic questions are "How much change has there been from the beginning of the program until now? At this rate of change, will we meet our objectives and goals by the end of the program period? What else is 'going on' about which we should be aware?"

c) Outcome measures will determine whether the objectives of the educational program have been met. These measures make it possible to summarize the progress made by the participants across the entire program. The basic questions are, "How much change did we effect this year? What do participants know now? Do they know what we had planned for them to know?"

An assessment system that includes all three of these key features, and leads directly to the evaluation, will provide. useful information for a variety of purposes, in a variety of modes, about a variety of participants. In other words, such a system will include multiple measures that provide information regardless of the participant's culture, gender, or language. Of course, it is assumed that the educational program will include valuable, worthwhile, and frequent opportunities to learn. Without the opportunity to learn meaningful material in a meaningful manner, an assessment system has little value. (As an example of a complete system of assessment, see Holt, 1994.)

Various types of assessments can, and should, be used within an appropriate assessment system. Each must be carefully thought out and be related to the others in some manner*. As a first layer of definition, an assessment may be *norm referenced, criterion referenced,* or may be an *alternative assessment* that describes current levels of knowledge, attitudes, and proficiencies. Some of the most frequently used are defined in Del Vecchio, et al., 1994.

Interviews and ***focus groups*** can provide in-depth information. In a structured interview, responses to a set of prepared questions can be recorded by the interviewer who can ask clarifying questions. Focus groups can include small groups of individuals and a facilitator to discuss a specific topic. Generally, scores are not developed; the data is qualitative in nature. It will be important to identify key individuals to interview (teachers, administrators, students, family members, and others in the community); it also will be important to create good questions to ask.

Surveys usually list a series of questions to be answered orally or in writing by the respondent. The responses can be *forced choice,* where the answers are provided (e.g., Are you pleased with the expertise of the staff facilitating the training sessions? yes/no), or may be scored on a rating scale (4 to 7 response options such as "very pleased with expertise" to "not at all pleased with expertise"). Scores can be developed by assigning point values to the responses (e.g., Yes=1, No=O) and summing these values. The responses also can be *open-ended,* where the individual provides an answer (e.g., What pleases you most about the expertise of the staff?). As with interviews, scores generally are not developed for open-ended surveys.

Observation checklists can be used to determine whether particular behavioral, physical, or environmental characteristics are present. Typically, desirable behaviors are described briefly and an observer checks (./) whether each behavior is observed during a particular period of time (e.g., the first week of the program). Scores can be developed by counting the number of checks. When the same checklist is used periodically throughout the program, it can be used to demonstrate progress by showing more behaviors being observed (checked) across time. In T-addition,, observational~ rating scales can.. be developed. To provide useful information, observational rating scales should be tied directly to the objectives and instructional activities of the program and conducted on a regular basis. By linking the descriptors and progression of ratings to instructional priorities, staff can obtain valuable data for assessing learners' ongoing progress and for improving the instructional program.

Alternative assessments are types of measures that fit a contextualized measurement approach. They can be easily incorporated into the training session routines and learning activities. Their results are indicative of the participant's performance on the skill or subject of interest. Observation measures are an example of an alternative assessment. As used within this document, "alternative assessment" subsumes authentic assessment, performance-based assessment, informal assessment, ecological assessment, curriculum-based measurement, and other similar forms that actively involve the participant.

For many types of alternative assessments, different scoring methods can be used. Three typically used methods are holistic scoring, which provides a general, overall score, *primary trait scoring,* which defines particular features (or traits) of a performance and then provides separate scores for each trait, and *analytic scoring,* which assigns a weight based on the importance of each trait (e.g., the use of inclusive language might be weighted more than correct grammar).

Criterion-referenced tests (CRTs) are sometimes considered as, a type of alternative assessment. CRTs measure whether specific knowledge has been gained; that knowledge being the criterion against which the participant's current knowledge is measured. Answers can be marked as correct or incorrect for scoring purposes. A score of 80% correct usually is considered as mastery of the knowledge.

Standardized tests can be used to measure participant skills. They are so named because their administration, format, content, language, and scoring procedures are the same for all participants -- these features have been "standardized." Locally developed and commercially available standardized tests- have been created for- most -achievement areas and for some aspects of language proficiency. When considering the definition of "standardized test," it is clear that all high-stakes tests should be standardized, whether they are commercially available tests or locally developed alternative assessments.

When referring to standardized tests, most people think of *norm-referenced tests* (NRTs). NRTs typically are used to sort people into groups based on their assumed skills in a particular area. They are useful when selecting participants for a particular program because they are designed to differentiate among test-takers. In addition, NRTs can provide general information that will help to match classrooms for overall achievement levels before assigning them to a particular program.

Portfolio does not refer to a specific type of assessment, but is an approach to organizing the information about an individual or a class/program. Portfolios can serve as a repository for "best" works or for all work on a particular project, from first notes to final draft. The portfolio can contain projects, assignments, various alternative assessments, and/or results from NRTs. The portfolio also can be used as a record of achievement that can be used to demonstrate expertise in a particular area.

E. Measures Relevant for Accountability to Specific Youngsters and Families

Below are listed a sample of promising instruments. It is essential that interveners review and choose measures that minimize negative impact on clients. Proper personalization of assessment in the best interests of the client may even call for not using a measure in its entirety or in the way the developer prescribes. We recognize that this violates standardization of administration and makes interpretation more difficult, but just as empirically supported therapeutic strategies must be adapted to ensure a good fit with a client, so must assessment practices. In both instances, empirical support for prevailing practices is not so strong as to warrant rigid implementation. Also of value are data from functional assessments (increasingly being done when students are referred for behavior problems). Finally, some interveners use projective procedures and selected items from other measures (e.g., sentence completion, drawings and related stories, Children's Depression Inventory) as a stimulus for discussion with clients. Client responses early and near the end of the period of intervention may be useful as supplementary evaluation data.

To find specific measures and read reviews about them, go to Buros Center for Testing online at http://marketplace.unl.edu/buros/ and search by type of measure, For example, use the following categories (we have listed a couple of examples of what you can find):

- **Client satisfaction (youngster; family)**
  >Youth Satisfaction Questionnaires (e.g., see
  http://www.dmh.cahwnet.gov/RPOD/child-posi.asp)
  >Vanderbilt Satisfaction Questionnaire (R.L. Oliver (1997). *Satisfaction: A Behavioral Perspective on the Consumer*, New York: Irwin/McGraw-Hill.

- **Reduction in Youngster's Symptoms/Problem Behaviors**
  >Child Behavior Checklist (Achenbach & Edelbrock; see
  http://buros.unl.edu/buros/jsp/reviews.jsp?item=13191584)
  >Child and Adolescent Functional Assessment Scale (Hodges; see
  http://buros.unl.edu/buros/jsp/reviews.jsp?item=06000977)

- **Increases in Positive Functioning**
  >Family Environment Scale (Moos; see
  http://www.docdatabase.net/more-family-environment-scale-third-edition-231902.html
  >Family Adaptability and Cohesion Scale (Olson; see
  http://facesiv.com/

- **Classroom/School Measures**

  Try this category on the Buros site to see the range of available meaures.

- **Systems of Care –** The following system-level measures are reviewed in *Measuring how we care: Tools for assessing children's mental health services, programs and systems* (1996) – prepared by Ted Cross, Matthew Urato, Heather Lyons, and Doreen Cavanaugh for the (former) Technical Assistance Center for the Evaluation of Children's Mental Health Services.

  >*The Survey of Parents' System of Care Experiences* (Epstein) – Parent/Guardian self-report

  >*Children and Youth Services Agency Network Survey* (Morrissey) – Agency representatives and mental health care providers within a service delivery system

  >*The Key Informant Survey Questionnaire* (Morrissey) – Agency staff representatives, consumer representatives, family groups, and/or government officials

  >*Local Children's Service System Interagency Collaboration Checklist* (Froelich) – Parents and representatives of agencies involved in the system

  >*Revised Interagency Cooperation Survey* (Moynihan) – One representative from each agency, as well as the community mental health center involved.

  >*System of Care Survey* (Kutash) – key informants, including parents, social workers, program administrators, etc.

Appendix

The Evaluation Problem

From: *On Understanding Intervention in Psychology and Education* by H. Adelman & L. Taylor (1994). Praeger. Online at http://smhp.psych.ucla.edu/pdfdocs/contedu/understandingintervention.pdf

> *Evaluation practiced at the highest level of the state-of-the-art is one means of speeding up the processes that contribute to human and social progress.*
>
> Rossi, Freeman, & Wright[1]

Increased concern about evaluation in psychology and education has advanced the way evaluation is conceived.[2] Despite the breadth of this scholarly activity, widespread demands for accountability continue to narrow the way professionals, clients, policy makers, underwriters, and the general public think about evaluation. Social and political forces literally have shaped the whole enterprise of program evaluation.[3]

The prevailing cry is for specific evidence of efficacy—usually in terms of readily measured immediate benefits—and for cost containment. Although understandable in light of the unfulfilled promise of so many programs and the insatiable demands on limited public finances, such naive accountability demands ignore the complexities of intervention. The problem is well exemplified by the narrow focus found in reviews, analyses, and reanalyses of data on psychotherapy, behavior change, and early education programs.[4]

Besides responding to accountability pressures, two unfounded presumptions are at the core of most current evaluations in psychology and education. One premise is that an intervention in widespread use must be at a relatively evolved stage of development and thus warrants the cost of summative evaluation. The other supposition is that major conceptual and methodological problems associated with evaluating intervention efficacy are resolved. The truth, of course, is that interventions are frequently introduced prior to adequate development with a view to evolving them based on what is learned each day. Moreover, many well-institutionalized approaches remain relatively underfunded and underdeveloped. As to the process of evaluation, every review of the literature outlines comprehensive, unresolved concerns. Given this state of affairs, accountability demands are often unreasonable and chronically reflect a naive view of research and theory.

Overemphasis on immediate evaluation of the efficacy of underdeveloped interventions draws resources and attention away from the type of intensive research programs necessary for advancing intervention knowledge and practice. Cost-effective outcomes cannot be achieved in the absence of cost-effective development of interventions and related intervention research. *Premature* efforts to carry out comprehensive summative evaluations clearly are not cost-effective. Consequently, policies mandating naive accountability run the risk of generating evaluative practices that are neither cost-effective nor wise.[5]

The evaluation problem, then, involves more than determining the efficacy of current interventions and more than finding better ways to evaluate efficacy. Broadly stated, it encompasses concerns about how to expand the focus of evaluation not only to contribute to improving practice, but also to aid in evolving theory and basic knowledge about intervention.

In the following sections, our intent is to briefly highlight (1) the concept of evaluation, (2) how the evaluation problem relates to the other three fundamental intervention problems, and (3) the key role of evaluative research in improving practice and advancing basic knowledge about intervention.

NOTES

1. P.H. Rossi, H.E. Freeman, & S. Wright (1979). *Evaluation: A systematic approach* (3rd ed.). Beverly Hills, CA: Sage.

2. For a comparison of evaluation models, see D.L. Stufflebeam & W.J. Webster (1983). An analysis of alternative approaches to evaluation. In G.F. Madaus, M.S. Scriven, & D.L. Stufflebeam (Eds.), *Evaluation models*. Boston: Kluwer-Nijhoff; also see P.H. Rossi & H.E. Freeman (1989). *Evaluation: A systematic approach* (4th ed.). Newbury Park, CA: Sage. E.J. Posavac & R.G. Carey (1989). *Program evaluation: Methods and case studies* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall. For recent reviews of the topic, see W.R. Shadish, Jr., T.D. Cook, & L.C. Leviton (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage. L. Sechrest & A.J. Figueredo (1993). Program evaluation. *Annual Review of Psychology, 44*, 645–674. M. Scriven (1993). *Hard-won lessons in program evaluation.* San Francisco: Jossey-Bass.

3. Recent reviews stress that the evolution of program evaluation in general and evaluation theory specifically has been shaped to a significant degree by evaluation researchers who were unprepared for their confrontations with complex social and political realities—including those associated with the shift from an industrial to a postindustrial (cybernetic) era. The demand for greater external validity has forced program evaluators to move beyond the prevailing paradigms and methods guiding the social sciences. For the most part, this demand reflects the socio-political-economic nature of intervention and evaluation. That is, interventions compete for limited societal resources and evaluation feeds into political decision making about which interventions are funded and levels of support.

4. See L. Bond & B.E. Compas (Eds.) (1989). *Primary prevention and promotion in the schools.* Newbury Park: Sage, pp.106–45). A. Kazdin (1990). Psychotherapy for children and adolescents. *Annual Review of Psychology, 41,* 21–54. M.J. Lambert, D.A. Shapiro, & A.E. Bergin (1986). The effectiveness of psychotherapy. In S.L. Garfield & A.E. Bergin (Eds.), *Handbook of psychotherapy and behavior change* (3rd ed.). New York: Wiley. A. Mitchell, M. Seligson, & F. Marx (l989). *Early childhood programs and the public schools: Promise and practice.* Dover, MA: Auburn House. R.E. Slavin, N.L. Karweit, & N.A. Madden (1989). *Effective programs for students at risk*. Boston: Allyn & Bacon. J.R. Weisz, B. Weiss, & G.R. Donnenberg (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist, 47*, 1578–1585.

5. Accountability pressures can lead to an overemphasis on immediate behavioral outcomes. Usually, decisions as to what and how to evaluate are made by those administering or funding an intervention. For example, with respect to specifying outcomes for evaluation, the primary focus in preparing IEPs for special education is on remedial outcomes. Furthermore, the prevailing emphasis is on specifying outcomes in terms of behavioral and criterion-referenced objectives. Similar trends are seen in psychology for interventions underwritten by third party payers. These trends no doubt are a major aid in efforts to evaluate whether outcomes are accomplished. However, the limited focus ignores the broader responsibility many interveners have for facilitating ongoing development and providing enrichment opportunities. A narrow focus on correcting problems also can be counterproductive to overcoming problems if the intervention involves little more than a set of laborious and deadening experiences. Moreover, many important facets of a program are not easily measured and thus may be given short shrift (e.g., self-concept, attitudes toward system improvement and problem solving). In general, the danger is that valuable intervention aims and goals are lost when *all* ends are specified in terms of highly concrete and easily measurable objectives. Not all complex long-range aims that an intervention should pursue can be stated as short-term or behavioral objectives. Indeed, only a relatively limited set of skills can be specified in highly concrete, behavioral terms—and even in these instances, it may not be desirable to do so for intervention purposes. In education, beside the fact that specifying everything in this way would result in far too many objectives to teach, the trend stresses *teaching* at the expense of *learning*. Moreover, attitudes, motivation, and creative functioning in the arts and sciences, for example, do not lend themselves to formulation in simple behavioral terms.

The dilemmas raised by accountability pressures are well illustrated in an article on mental health services for children: see J.D. Burchard & M. Schaefer (1992). Improving accountability in a service delivery system in children's mental health. *Clinical Psychology Review, 12*, 867–882.

1

The Essence of Evaluation

When the cook tastes the soup it is formative evaluation and when the guests taste the soup it is summative.

Stake[1]

Evaluation involves determining the worth or value of something.[2] In formal terms, we define comprehensive evaluation as a systematic process designed to describe and judge an intervention's antecedents, transactions, and overall impact and value for purposes of making decisions and advancing knowledge.[3]

Everyone evaluates interventions with which they come in contact. Whenever anyone decides that an intervention is or isn't a good one, an evaluation is made.[4] Interveners judge whether their own and others' programs are going well. Clients are quick to formulate likes or dislikes of interveners and their programs. Administrators know which programs they think are working and which aren't.

Some evaluative judgments simply reflect an individual's or group's informal observations. Other judgments are based on careful data gathering and analyses and use of appropriate sets of standards. Some evaluations only offer conclusions about the degree to which a program is effective. Most, however, also incorporate the conclusions of those judging the program in terms of whether they agree with what it is trying to do. Since what a program intends to do stems from its rationale, program evaluations inevitably influence views about the appropriateness of its underlying rationale.

Systematic evaluation planning requires decisions about (1) the focus of evaluation (e.g., person or environment, immediate objectives vs. long-range aims), (2) whose perspective (e.g., client, intervener, program underwriter) is to determine the evaluation focus, methods, and standards used, and (3) the best way to proceed in gathering, analyzing, and interpreting information (e.g., specific measures, design). In making such decisions, concerns arise because what can be evaluated currently is far less than what a program may intend to accomplish. Furthermore, inappropriate bias and vested interests shape evaluation planning and implementation, thereby influencing whether a program is seen as good or bad. And all aspects of evaluation have the potential to produce negative effects; for instance, evaluation can lead to invasion of privacy and an undermining of the ability of clients and interveners to self-evaluate, and over time, what is evaluated can reduce and reshape a program's intended aims.

PURPOSES

Intervention evaluation can aid efforts to (1) *make decisions* about whether to undertake, continue, modify, or stop an intervention for one or more clients and (2) *advance knowledge* about interventions in ways that can advance understanding of and improve practices (including utility), training, and theory. Evaluation is useful in relation to a great variety of interventions as an aid in assessing efficiency, effectiveness, and impact. As Rossi and Freeman state:

> The mass communication and advertising industries use fundamentally the same approaches in developing media programs and marketing products; commercial and industrial corporations evaluate the procedures they use in selecting and promoting employees and organizing their work forces; political candidates develop their campaigns by evaluating the voter appeal of different strategies; . . . administrators in both the public and private sectors are continually assessing clerical, fiscal, and interpersonal practices of their organizations. The distinction between these uses of evaluation lies primarily in the intent of the effort to be evaluated . . . to benefit the human condition . . . [or] for other purposes, such as increasing profits or amassing influence and power.[5]

Providing a broad categorical view of the areas in which evaluation is applied, Scriven outlines the "Big Six" plus others. The Big Six are listed as product, performance, personnel, program, proposal, and policy evaluations. To these, he adds two other applied fields. "The first is the evaluation of evaluations (meta-evaluation). . . . The second is a field comprising a set of fields: It might be called `intradisciplinary evaluation,' the evaluation of the data, sources, explanations, definitions, classifications, theories, designs, predictions, contributors, journals, and so on within a discipline." Scriven concludes: "In toto, intradisciplinary evaluation is by far the largest part of evaluation, and having practitioners do it with reasonable skills is the price of admission to the company of disciplines. Other applied fields besides the Big Six range from literary criticism and real estate appraisal to quality control in industry."[6]

Stake's evaluation matrix is reproduced in Figure 1 as an example of a framework designed to outline the general nature of information for meeting many evaluation purposes.[7] As the framework suggests, evaluation encompasses the acts of *describing* and *judging* an intervention's (1) rationale, including assumptions and intentions, (2) standards for making judgments, (3) actual activity, including intended and unintended procedures and outcomes, and (4) costs—financial, negative effects, and so forth. To achieve the above ends in a comprehensive manner, both immediate and long-term information on an intervention must be gathered.[8]

TASKS FOR PLANNING

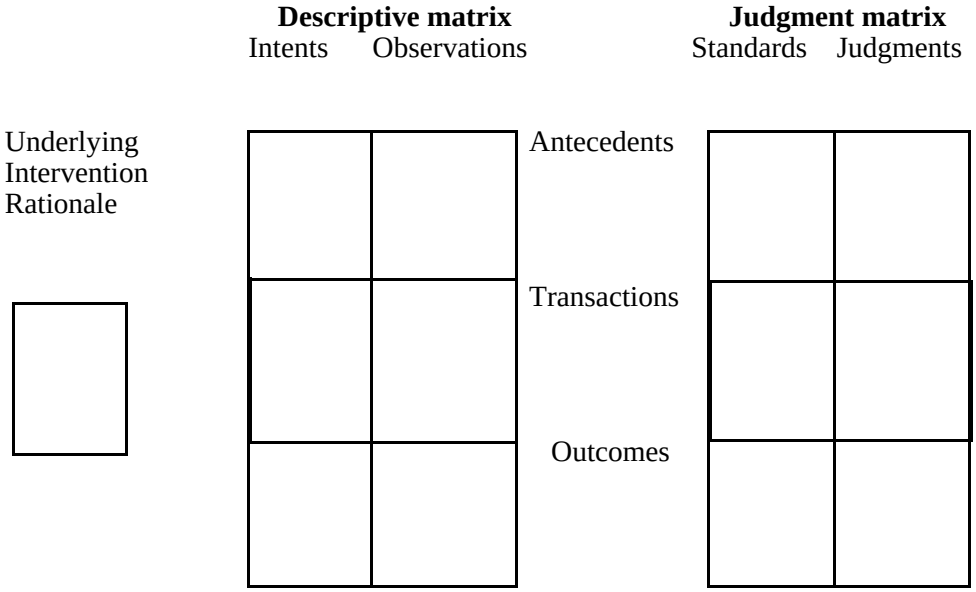
Awareness of tasks involved in planning an evaluation provides another perspective on the process. Such tasks reflect the necessity in evaluation planning of making decisions about the focus of the evaluation, its specific objectives, and appropriate methodology and measures.

Our formulation identifies the following seven key planning tasks:

- *Clarifying the intended use of information.* Most important here is awareness of who wants the information and why they need it. Ultimately this translates into the question: What types of decisions are to be made? Also important is the matter of anticipating the use and political and motivational impact of evaluation processes and findings. This includes a significant appreciation of the often conflicting interests among the variety of interested parties (i.e., stakeholders).
- *Understanding the intervention's rationale.* In cases where evaluation includes judging the intervention rationale, pursuit of the above task (clarifying the intended use of evaluation information) will result in gathering information about the rationale. However, when the evaluation is designed with reference to a standardized set of objectives, clarification of the rationale becomes a separate task. In either case, an understanding of the intervention rationale can provide a separate basis for deciding about other intervention facets to evaluate.

Figure 1

Layout of Statements and Data to Be Collected During Evaluation



Source: R. Stake (1967). The countenance of educational evaluation. *Teachers College Record, 68,* 523–40. Reprinted with permission.

- *Formulating evaluation questions.* Evaluative concerns are translated into a set of questions. For example: Were intended antecedent conditions present during the intervention? Which procedures were effective for which clients? Were there undesirable transactions? Were specific objectives achieved? Were long-range aims achieved? Did expected negative outcomes occur? Were there unexpected negative outcomes? •*Specifying information to be gathered.* Relevant descriptive information that can answer each major question is specified. The more things one is interested in evaluating, the more one has to settle for samples of information. Some of the information likely will be of a quantitative nature; some may be qualitative.[9]
- *Specifying procedures.* Decisions about information gathering are shaped first by what one wants to know and then are tempered by practical considerations. Problems related to gathering desired information become evident as one attempts to specify procedures. Limitations related to time, money, sample availability, valid measures, multivariate statistics, and personnel usually lead to major compromises in evaluation planning. For example, sometimes a good measuring instrument exists; sometimes only weak procedures are available; sometimes gathering desired information is not currently feasible. A special set of problems stems from the socio-political-economic concerns (e.g., threats to current status) and psychological reactance (e.g., fear-based resistance) that are common phenomena when evaluation is introduced.[10]
- *Specifying a design.* An evaluation design is used so that information can be gathered and interpreted appropriately. When someone asks how good an intervention is, judgments are based on the available information and are relative to some standard of comparison. A sound design ensures that appropriate bits of information (e.g., data) are gathered, including information for use as standards for judgments. A sound evaluation design also includes provision for the gathering and use of information for revising interventions as the process proceeds.[11]
- *Designating time and place for collecting information.* Further practical considerations arise when evaluations are scheduled. The design sets the general parameters; the particulars are determined by practical factors such as resource availability.

One major evaluation concern not reflected above involves decisions about the role of various interested parties. For example, as suggested throughout, rationales may differ with respect to what should be evaluated. If so, whose rationale should prevail? Every facet of an evaluation is influenced by the answer to this question.

Another matter not specifically addressed above involves ethical concerns associated with evaluation. Naturally, these are similar to those discussed in relation to assessment in general. For instance, evaluators must be concerned with how to minimize possible bias and conflicts of interest, as well as negative consequences that can arise from evaluation itself.

IMPACT ON PROGRAM BREADTH

As the discussion to this point underscores, a common use of evaluation is to determine if one agrees with what the intervention is trying to accomplish and how well the intervention is accomplishing the full range of outcomes desired. The less a program is trying to achieve, the easier it is to determine these matters. It is hard to evaluate large-scale social programs, community agencies, and most school programs, for example, because they are trying to accomplish so many different goals.[12]

Ironically, the longer a program is subjected to external, formal evaluation, the less it may try to accomplish. At least this seems to be one negative effect of the big push toward behavioral and criterion-referenced outcomes as ways to improve accountability. That is, such approaches can cause a shift away from a program's long-range aims toward a limited set of immediately measurable objectives. This is a negative form of "teaching to the test" because, in the process, many important things are ignored simply because they will not be directly evaluated.[13] If one is not careful, the desire for information on effectiveness can redesign a program's underlying rationale in ways that inappropriately reduce its breadth of focus.

Comprehensive evaluation should stress the full scope of desired intervention aims. That is, even when certain processes and outcomes are not easily measured, they still must be evaluated as well as is possible and kept in the forefront of discussions about a program's worth. For example: from a motivational perspective, a basic concern is whether a program enhances clients' interest, desire, and participation in improving their functioning. Because none of these outcomes is readily measured, the danger is that they will not be afforded the attention they warrant.

In sum, evaluations of whether an intervention is any good must first address the question: Is what it is trying to accomplish appropriate? The frame of reference for such evaluations may be the intervention rationale or what others think the program should be doing or both. After judging the appropriateness of what is wanted or expected, a program's intended breadth of focus should guide efforts to evaluate effectiveness. Because not everything is measurable in a technically sophisticated way, some things will be poorly measured or simply reviewed informally. Obviously, this is less than satisfactory. Still, from a rational perspective, continued emphasis on the entire gamut of what is intended is better than limiting evaluation to approaches that inappropriately narrow the breadth of focus for intervention.[14]

In this context, we are reminded of Yankelovich's commentary on measurement:

> The first step is to measure whatever can be easily measured. This is okay as far as it goes. The second step is to disregard that which can't be measured or give it an arbitrary quantitative value. This is artificial and misleading. The third step is to presume that what can't be measured easily isn't very important. This is blindness. The fourth step is to say what can't be measured really doesn't exist. This is suicide.[15]

NOTES

1. R.E. Stake (1976). *Evaluating educational programs: The need and the response.* Paris: Organization for Economic Cooperation and Development, p. 19.

2. We recognize the deficiencies of this simple definition. Still, it conveys the essence of the process. Reviewing the matter, Scriven states: "Evaluation is a process of determining certain evaluable properties of things, but there is more than one kind of such properties. Perhaps the most fundamental and important distinction among them is between merit or quality and worth or value." Using the example of a high school French teacher, he notes that the teacher may be the best in a school, but if enrollment patterns shift away from French, that teacher's worth or value to the school diminishes. The teacher's merit (i.e., quality in terms of professional standards) has not declined, but his or her benefit (vis à vis meeting the school's needs) has. M. Scriven (1993). *Hard-won lessons in program evaluation.* San Francisco: Jossey-Bass, p. 67.

3. Rossi and Freeman use the terms *evaluation* and *evaluation research* interchangeably. Their definition states: "Evaluation research is the systematic application of social research procedures for assessing the conceptualization, design, implementation, and utility of social intervention programs." See P.H. Rossi & H.E. Freeman (1989). *Evaluation: A systematic approach* (4th ed.). Newbury Park, CA: Sage, p. 18.

4. Conclusions of good or bad clearly are value judgments. Shadish and colleagues note that "Early evaluators mostly ignored the role of values in evaluation—whether in terms of justice, equality, liberty, human rights, or anything else. . . . such evaluators believed their activities could and should be value-free. But it proved to be impossible in the political world of social programming to evaluate without values becoming salient. Social programs are themselves not value-free." W.R. Shadish, Jr., T.D. Cook, & L.C. Leviton (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage, pp. 46–47.

5. Rossi & Freeman, *Evaluation*, p. 19.

6. Scriven, *Hard-won lessons in program evaluation*, p. 44.

7. R.E. Stake (1967). The countenance of educational evaluation. *Teachers College Record, 68,* 523–40. Among program evaluators, Robert Stake is one of the early and long-term contributors. See Shadish, Cook, & Leviton, *Foundations of program evaluation*, for a comprehensive overview of his ideas and contribution, as well as those of other influential leaders such as Michael Scriven, Donald Campbell, Carol Weiss, Joseph Wholey, Lee Cronbach, and Peter Rossi.

8. A relatively new form of evaluation practice is a process called "prospective evaluation," which has been developed by the Program Evaluation and Methodology Division (PEMD) of the U.S. General Accounting Office (GAO). The purpose of the process is to predict or forecast the impact of a proposed program or policy change (e.g., as an aid to legislators). The potential value of such forecasts is obvious; so are the problems associated with efforts to make accurate predictions. See General Accounting Office (1989). *Prospective evaluation methods: The prospective evaluation synthesis*. GAO/PEMD-89-10. Washington, DC: Author.

9. Among academics, there is a running argument about the relative merits of quantitative and qualitative evaluations. In response to the many who argue primarily for quantitative evaluation, Guba and Lincoln have argued strongly for qualitative evaluation. See E.G. Guba & Y.S. Lincoln (1989). *Fourth generation evaluation.* Newbury Park, CA: Sage.

Sechrest and Figueredo suggest that a compromise may be possible "in light of the realization that although rigorous theory testing is admittedly sterile and nonproductive without adequate theory development, creative theory construction is ultimately pointless without scientific verification." L. Sechrest & A.J. Figueredo (1993). Program evaluation. *Annual Review of Psychology, 44,* 645–74, p. 654.

10. Posavac and Carey enumerate and discuss how political and psychological factors can undermine evaluation efforts, and suggest ways to plan for dealing with them. See E.J. Posavac & R.G. Carey (1989). *Program evaluation: Methods and case studies* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

11. Tharp and Gallimore describe a fine example of program development based on a progressive series of formative and summative evaluations. Over a period of ten years, they made a succession of process and outcome evaluations using the quantitative data and qualitative information gathered on variables affecting the outcomes to improve the program. That is, data gathered at each stage of program development were used as feedback for revising the intervention. See R.G. Tharp & R. Gallimore (1979). The ecology of program research and evaluation: A model for evaluation succession. In L. Sechrest, S.G. West, M.A. Phillips, R. Redner, & W. Yeaton (Eds.), *Evaluation Studies Review Annual* (Vol. 4, pp. 39–60). Beverly Hills, CA: Sage.

12. Besides being difficult to carry out, evaluations of large-scale social and educational programs are costly, and the history of efforts to evaluate such programs is characterized by weak and often poorly conceived methodology as well as findings that are subject to varying interpretations. At the same time, it is evident that such evaluations must be pursued, and we must learn to do them better. In this regard, each new national and state evaluation provides a unique opportunity to improve the process of evaluation.

13. Charles Silberman cogently noted in his 1970 book, *Crisis in the classroom* (Vintage Books): "Elementary school students almost invariably regard mathematics as the most important subject in the curriculum—not because of its elegance, but because math has the most homework, because the homework is corrected the most promptly, and because tests are given more frequently than in any other subject. The youngsters regard spelling as the next most important subject, because of the frequency of spelling tests" (p. 147).

We would add that, with increasing demands for accountability, teachers quickly learn what is evaluated and

what is not, and slowly but surely greater attention is given to teaching what will be on the tests. Over time, what is on the tests becomes viewed as what is most important. Because only so much time is available to the teacher, other things not only are deemphasized, they also are dropped from the curriculum. If allowed to do so, accountability procedures have the power to reshape the entire curriculum.

What's wrong with that? Nothing—if what is evaluated reflects everything we want students to learn in school. Unfortunately, this is not the case.

Current accountability pressures reflect values and biases that lead to evaluating a small range of basic skills and doing so in a narrow way. For students diagnosed with problems, this is seen in the fact that their school programs increasingly have been restricted to improving skills they lack. As a result, they are cut off from participating in learning activities that might increase their interest in overcoming their problems and that might open up opportunities and enrich their future lives.

14. The issues related to the impact of a narrow focus on evaluation also arise in the context of discussions about evaluating intervener competence. That is, narrowly focused competency evaluations may constrict rather than expand intervener growth with respect to the broad range of knowledge, skills, and attitudes needed to properly plan, implement, and evaluate interventions.

15. Cited in A. Smith. *Supermoney*. New York: Random House, p. 286.

2

Evaluation and the Other Basic Intervention Problems

In the last analysis, we see only what we are ready to see. We eliminate and ignore everything that is not part of our prejudices.

Charcot

The complexity of evaluation is best appreciated when viewed in the context of the other three fundamental intervention problems (see Figure I.1 in the Introduction). That is, each problem has implications for evaluation. Reciprocally, research of an evaluative nature is essential in advancing knowledge about these intertwined problems; each must be reasonably well addressed before interventions can be optimized. For instance, appropriate decisions about processes and outcomes cannot be made logically in the absence of valid differentiation among systems with regard to intervention needs. If appropriate process and outcome decisions cannot be made, appropriate planning and implementation are jeopardized. Thus, even when valid evaluation practices are available and feasible, the logical prerequisites for a sound summative evaluation of efficacy may be absent. The following discussion highlights a few basic implications for evaluation of the classification, underlying rationale, and planning and implementation problems.

CLASSIFICATION

Because the most used classification schemes in psychology and education focus on person problems, it is particularly instructive to look at evaluation in that context. As discussed in Part I, common approaches to classifying psychological and educational maladies categorize them in terms of current dysfunctioning, causal factors, prescriptive implications, or some combination of all these. The variables and criteria used in defining a category usually are chosen because they have immediate relevance for research, intervention, administrative, or policy matters.

When the emphasis is on current dysfunctioning, both severity and pervasiveness of dysfunction are relevant concerns. That is, a narrow or broad range of areas of human functioning may be affected, such as one or more developmental areas or facets of school, family, or job performance. Additionally, severity of dysfunction may range from mild to profound, depending on normative expectations related to factors such as levels of development and competence and socioeconomic and subcultural status. As graphically suggested in Figure 2.1 (see Part I), the combination of pervasiveness and severity yields nine classification groups when treated as discrete categories rather than continuous variables. When the paradigmatic cause of the dysfunction is added as a third dimension, the schema jumps to twenty-seven groups. Adding duration results in an another exponential increase in categories.

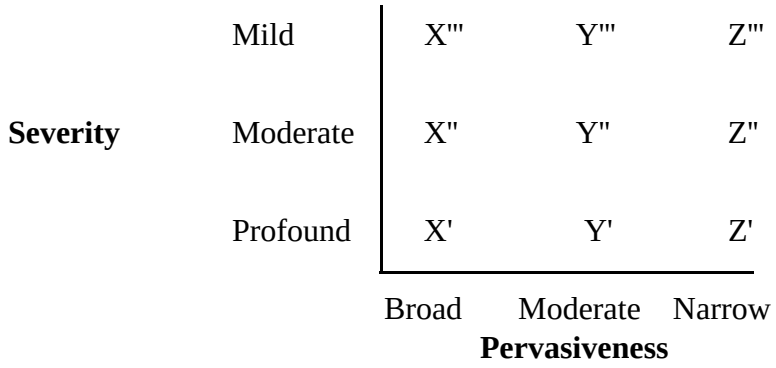
One clear implication of the relatively simple nine-group classification (see Figure 2) is that outcomes are likely to be easier to achieve with the Z'" category than for those in the X' group. With regard to efficacy, then, groups *minimally* should be separated for evaluation based on degrees of pervasiveness and severity of dysfunctioning at the onset of intervention.

Classification in terms of cause demonstrates other complexities. At times, causal factors and their effects logically are key indicators for intervention decisions and are potential predictors of outcome. In these cases, classification based on causal factors and their current manifestations are of great significance in evaluating intervention. Such categorization can be done using primary instigators, secondary contributing factors, or both. (The situation becomes extremely complex once

secondary factors begin interacting with primary instigators.) The type of primary instigator and the degree of dysfunction produced by it determines the degree to which secondary factors exacerbate problems. In Part I, we pointed to the example of physiological "insult" causing a major CNS disorder, the effect of which is so severe and pervasive that result ant dysfunctioning cannot be significantly worsened. We contrasted this with cases of minor CNS disorders where a great many secondary variables can aggravate existing dysfunctions and create other problems. Outcomes differ for such contrasting groups and for groups that vary in the degree to which the pathological impact of causal factors can be compensated for or reversed. Therefore, such group variations should be accounted for in evaluating interventions.

Figure 2

Evaluation with Reference to Classification of the Severity and Pervasiveness of Dysfunctioning



Failure to identify appropriate subgroups is a key factor in controversies over intervention efficacy. Examples are plentiful: the Head Start program evaluations, psychotherapy outcome studies, the evaluation of alcoholism treatments, investigations of learning disability interventions. In most cases, inadequate efforts are made to discriminate between important subgroups in analyzing findings, and as a result, premature conclusions are reached about efficacy or lack thereof.

Researchers who focus on the intervention-client match have stressed that the question is not: Does an intervention work? The proper question is, To what degree does a specific intervention work for a given subgroup? In evaluating efficacy, classifications are used to specify relevant subgroups that can help clarify sources of variance in process and outcome data. Analyses of evaluation findings are shaped in basic ways, then, by work done on the classification problem.

UNDERLYING RATIONALE

Concerns about the quality of underlying rationales and factors that bias the formulation of such rationales highlight the importance of initiating evaluation processes *prior* to the onset of intervention. For instance, evaluations can end up reifying biased rationales and undermining alternative points of view. The common example in psychology and education arises when a rationale inappropriately designates individuals as the problem, thereby blaming the victim. Under

such circumstances, efforts to correct the problem and measure outcomes primarily focus on individuals. Ignored are alternative views of cause and correction, such as the possibility that the environment is the source of the problem and the proper focus for intervention.

When alternatives are not considered, prevailing biases tend to go unchallenged. Such consequences are minimized through preimplementation evaluations of intervention rationales and plans. These allow one to judge a priori the appropriateness of decisions about who or what will be the focus of intervention and about the nature and scope of outcome objectives. Furthermore, given the importance of the rational relationship between means and ends, a priori judgments can be made about the compatibility between procedures and projected outcomes. (This is the type of predictive process the U.S. General Accounting Office calls prospective evaluation and uses to forecast the impact of a proposed program or policy change as an aid to legislators.) Also, well before the time comes for analyzing outcomes, decisions can be made about whose objectives will be weighted most heavily in case of conflicting interests.

In general, evaluative investigations of the utility and validity of intervention rationales can have an immediate impact on practice, for example by minimizing the perpetuation of systematic biases and the implementation of interventions that are likely to produce more harm than good. And such research also should have a cumulative impact on basic knowledge and theory.

PLANNING AND IMPLEMENTATION

Implications of the planning and implementation problem for evaluating efficacy are so straightforward as to seem simpleminded. Sound planning calls for a logical consistency among intended antecedents, processes, and outcomes. Desired outcomes are unlikely when antecedent conditions necessary for intervention success are significantly lacking or when planned processes are not a logical match with antecedents. As suggested above, one basic evaluation implication is that intervention plans should be judged prior to implementation. Such judgments decide (1) whether there is a logical consistency among intended antecedents, processes, and outcomes, (2) likely costs and benefits, (3) the degree to which intended antecedents are present, and (4) whether observed antecedents are congruent with planned processes.

With respect to implementation, processes should be monitored from the moment they are initiated to determine if intended transactions occur and to detect unintended interfering transactions. Such information enables comparisons of optimal, average, and unsatisfactory implementations. For instance, when intended antecedents and transactions are not well approximated, desired outcomes are unlikely. An extreme but commonplace example is seen when a client terminates an intervention prematurely because of factors such as poor motivation or financial difficulty. By definition, an intervention is not optimally implemented in such cases. Consequently, outcome information on these clients obviously should be differentiated from findings on those for whom the intended antecedents and processes were logically consistent and carried out as planned.

A comparable example arises when clients stay but are not motivated to cooperate with interveners. Sound planning, implementation, and efficacy evaluation calls for addressing motivational differences among clients. This highlights another focus for classification. That is, clients should be differentiated in terms of degree of motivation (high, moderate, low) and its valence (positive, negative). This example, once again, underscores the interrelationship among intervention problems.

3

Studying Intervention—Not Just Evaluating Efficacy

Do not follow where the path may lead. Go instead where there is no path and leave a trail.
Anonymous

As the preceding presentation stresses, discussions of the evaluation problem generally deal with how to maximize the quality of information on intended outcomes. To a lesser degree, the literature also highlights the significance of gathering information on unintended outcomes, especially negative side effects, in evaluating costs versus benefits.[1]

Because the primary focus is on evaluating efficacy, the critical importance of evaluating other facets of intervention is less appreciated. Analyses of the evaluation problem can be instrumental in altering this situation, especially when bolstered by implications derived from work on the other three fundamental intervention problems. Such analyses underscore that, not only should other facets be evaluated, but at times these facets should be the primary and even the only focus of intervention evaluation.[2] Indeed, this must happen in order to improve the evaluation of programs *and* advance basic knowledge about intervention as a general phenomenon.[3]

IMPROVING INTERVENTION EVALUATION

Obviously, practitioners should improve their interventions and be accountable. Equally evident is the need to upgrade the methodological and technological infrastructure for evaluation.

We have stressed that evaluations can as easily reshape interventions in negative as in positive directions.[4] As an example, we have described how some accountability practices can inappropriately narrow a program's focus. This is seen in programs that limit evaluation to specific treatment and remedial objectives and overemphasize processes for pursuing these outcomes at the expense of those that foster ongoing client development in unaffected areas.

Finding out if an intervention is any good is a necessity. But in doing so, evaluation is not simply a technical process. The processes involve decisions about what and how to measure, and these decisions are based in great part on values and beliefs. As a result, limited knowledge, bias, vested interests, and ethical issues constantly influence the descriptive and judgmental processes and shape the decisions made at the end of the evaluation. Ultimately, the decisions made affect not only individuals but the entire society.

For new and evolving major interventions, evaluation must be extremely broad and used for formative purposes. Expending limited resources on comprehensive summative evaluations in the early stages of an intervention's development usually is premature. Resources are better devoted to formative evaluations. For instance, the activity might encompass the type of programmatic research and development required to test and improve unvalidated, large-scale treatments. Even for an established intervention, more than outcome information is necessary for describing and judging an approach's ongoing impact and value.

As we have also emphasized, preimplementation evaluations are essential for making judgments about whether a plan is worth carrying out. To fill out our preceding discussion just a bit: Evaluation can answer many questions before an intervention is implemented. One such question

is whether the underlying intervention rationale is coherent, logical, and well grounded theoretically and empirically. If it is not, there may be little justification for proceeding. Similarly, prior to implementation, determination should be made of how well key variations in antecedent conditions are addressed. For example, do the intended processes account for existing individual and subgroup differences? That is, do they address differences in the severity and pervasiveness of problems, differences in motivation for overcoming problems, and so forth? When such variations cannot be validly classified, subsequent efforts to judge the impact of specific interventions often are futile.

Given the presence of relevant antecedent conditions, key evaluative concerns include whether intended processes actually occur and whether potent unintended processes transpire. Findings related to such matters are basic in deciding whether to evaluate efficacy and how to interpret assessed outcomes. Also basic to such deliberations are conclusions about whether projected outcomes and available measures are proper indicators of efficacy. When they are not, outcome evaluation probably won't be productive, and the resources would be better used to refine the intervention or validate evaluation measures.

ADVANCING INTERVENTION KNOWLEDGE

Beyond expanding awareness about the nature and scope of intervention evaluation, understanding the evaluation problem underscores the role of evaluative research in advancing basic knowledge about interventions. Such activity is essential if intervention practices are to meet society's needs and expectations. As Campbell has suggested, evaluative research can be part of an experimental approach to social reform "in which we try out new programs designed to cure specific social problems, in which we learn whether or not those programs are effective, and in which we retain, initiate, modify, or discard them on the basis of apparent effectiveness on the multiple imperfect criteria available."[5] Obviously, knowledge-driven research and decision-driven research are not necessarily incompatible. The danger in both cases arises when the fuel supply (financial support) is scarce and its distribution is tied to naive or biased accountability practices.

On a theoretical level, evaluation research is advancing knowledge about classes of intervention (e.g., community, organizational, educational, mental health) and about intervention as a pervasive phenomenon. Increasingly, such research is guided by and contributes to model building about the general nature of intervention. Initial efforts to evolve a comprehensive model of intervention use frameworks and concepts as aids to investigating significant commonalities and differences among interventions. This work also shows promise for stimulating more comprehensive and systematic research and theory designed to improve cross-intervention understanding of what works and what does not, and why.

Examples of the types of questions for which answers are beginning to emerge are: What is the essence of intervention? Are there specific philosophical and theoretical concepts that underlie all intervention activity? What are the major elements found in all interventions? These and other questions raised throughout this monograph represent researchable topics that can contribute to fundamental knowledge and theoretical thought regarding the phenomenon of intervention.

NOTES

1. For a prototypical illustration of the importance of gathering data on unintended outcomes, see G.L. Barkdoll (1992). Strong medicine and unintended consequences. *Evaluation Practice, 13*, 53–57

2. Shadish and colleagues point out that intervention evaluation contributes to understanding related to (1) social programming (e.g., how programs and policies develop), (2) knowledge construction (e.g., how we learn about social action), (3) valuing (e.g., how interventions are judged), (4) knowledge use (e.g., how available knowledge is used as rationales for intervention), and (5) evaluation practice (e.g., how evaluations are conducted and improved). See W.R. Shadish, Jr., T.D. Cook, & L.C. Leviton (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.

On evaluative research, also see H. Chen & P. Rossi (Eds.) (1992). *Theory-driven evaluations in analyzing policies and programs*. Westport, CT: Greenwood Press. R.T. Gottfredson (1984). A theory-ridden approach to program evaluation: A method for stimulating researcher-implementer collaboration. *American Psychologist, 39*, 1101-12. J.C. Masters (1984). Psychology, research, and social policy. *American Psychologist, 39*, 851–62. R.G. Tharp & R. Gallimore (1979). The ecology of program research and evaluation: A model of evaluation succession. In L. Sechrest et al. (Eds.), *Evaluation Studies Review Annual* (Vol. 4). Beverly Hills, CA: Sage.

Note also that professionals concerned about advancing the state of the art related to evaluation are organized in the United States as the American Evaluation Association and in Canada as the Canadian Evaluation Association. The two groups plan to hold the First International Congress on Evaluation in 1995.

3. As Sechrest and Figueredo note, extending evaluation into new areas will help differentiate generic from specific concerns. See L. Sechrest & A.J. Figueredo (1993). Program evaluation. *Annual Review of Psychology, 44*, 645–74.

4. A much neglected area for research is the *psychology of evaluation*. From observation and personal experience, most of us know that anticipating and experiencing evaluation produces major reactive effects. Systematic studies are needed of the prospective, in-process, and antecedent psychological impact of evaluation on (1) systems that are evaluated, (2) evaluators, and (3) those who use evaluation findings.

5. D.J. Campbell (1969). Reforms as experiments. *American Psychologist, 24*, 409–29, p. 409.

A Few Concluding Comments about the Evaluation Problem

No longer is it assumed that well-meaning individuals or groups . . . [responsible for] health, education, training, rehabilitation, or other service actually help.

Posavac & Carey[1]

Intervention evaluation is difficult, and a task many would prefer to avoid.[2] This is part of the evaluation problem.

Everyone agrees that practitioners should be accountable, but there are major disagreements about what that means. Obviously, practitioners must show that their work is effective. But effective in what way? To what degree? At what cost? These questions underscore another aspect of the evaluation problem.

In choosing what we look at, how we observe, what we perceive, and what we report, interveners are strongly influenced by society's values, policies, priorities, and rewards. These influences, of course, usually are mediated by the predilections of those who employ us and by our personal and professional codes of ethics and values, favored models, and so forth. When one doesn't agree with an intervention's rationale, one will not likely approve of the intervention, even if evaluation findings indicate that it is effective. These also are facets of the evaluation problem.

Methodologically, evaluation must be carried out with exceedingly limited tools. So technical limitations add significantly to the problem.

And almost everyone has experienced negative consequences from evaluation. Those evaluated often are harmed, and consumers of evaluation reports frequently are misled. Evaluations create tensions and dilemmas and can be misused to create undesirable degrees of uniformity and conformity. Ultimately, we should be as concerned with the consequences of evaluation as we are with improving our technical capability to conduct better evaluations.

All this said, perhaps the biggest problem related to intervention evaluations is that they rarely are designed in ways that truly improve programs and advance knowledge.

Evaluation is the door to the future. Ironically, prevailing evaluations of intervention seem to be closing rather than opening that door.

NOTES

1. E.J. Posavac & R.G. Carey (1989). *Program evaluation: Methods and case studies* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall, p. 3.

2. On the topic of what types of difficulties to anticipate, see M.S. Scriven (1993). *Hard-won lessons in program evaluation*. San Francisco: Jossey-Bass.