BEYOND THE NUMBERS:
Labor Market Information Research and Writings

An Occasional Paper Series Published by Texas State Occupational Information Coordinating Committee
3520 Executive Center Drive, Suite 205, Austin, Texas 78731-1637 (512) 502-3750

June 1, 1997 Number 4

Roles and Responsibilities in a Performance Measurement System: Description, Prescription, and Policy-Making

by Marc Anderberg and Richard Froeschle

Any education and workforce development system must have a feedback mechanism to determine participant outcomes and program performance if it is to improve services based on the needs of customers and adapt to changing labor market conditions. To implement a systems approach to performance measurement, consensus must be achieved on activities and procedures as well as the roles and responsibilities of key stakeholders. The more fragmented the education and workforce development programs are at the outset, the more difficult it is to bind all the stakeholders to common definitions, measures, standards, adjustment models, and uniform data collection methods.

In Texas, the process of building a comprehensive performance measurement system has been underway since 1989 when the first feasibility study for using Unemployment Insurance (UI) wage records to identify student outcomes was published. Since that time, many hours of confrontation, conflict, and conciliation have passed to a point where there is widespread agreement on a set of core performance measures, key program definitions, and database file structures capable of providing the underpinnings of a comprehensive system.

This performance measurement and evaluation system has many complex technical, definitional, and political features that must be revisited and refined constantly. The partner agency and legislative stakeholders have changed; the political and programmatic environments have been altered. Despite significant changes in the external environment, steady progress has been made in developing and institutionalizing Texas's comprehensive performance measurement system in three major areas: (1) **data collection**; (2) **program evaluation**; and (3) **policy response**.

Data Collection

Data collection provides the raw numbers that serve as the basis of program evaluation. Without the collection of appropriate and reliable data, and the organization and reporting of data under designated performance measures, program evaluation becomes little more than anecdotal conjecture. Moreover, there is no value in establishing interesting and politically palatable performance measures when the necessary data collection is neither technically feasible nor cost effective.

Data collection includes:

- devising and agreeing to operational definitions of each **common program measure**;
- securing required interagency **data exchange agreements**;
- establishing **procedures for data coding and storage** which facilitate efficient data **transmission** and **retrieval** while safeguarding **privacy and confidentiality** of individually-identifiable and firm-specific information;

• implementing a **valid, reliable, and cost-effective methodology** for gathering necessary data;

• generating frequency distributions, cross-tabulations and other **descriptive statistics** which tell how many former students and program participants achieved specific outcomes under what conditions;

• expanding the breadth of options to approach 100 percent of all possible outcomes.

Program Evaluation

By means of contrast, **program evaluation** goes beyond descriptive statistics. Program evaluation requires **establishing a framework for setting program benchmarks and performance standards** which are used to classify programs based on hard evidence about their outcomes. This component includes policy simulations; i.e., "what-if" analysis. Policy simulations, for example, ask questions such as: "*What* percentage of programs would exceed or fail *if* the standard for post-exit employment is set at 75 percent?" or "*What if* the standard for earnings at entered employment is raised by $100 per quarter?". In addition to running simulations to set reasonable standards, program evaluation includes **creating and applying an adjustment model** to account for varying sub-populations served, regional economic conditions etc. so that all programs may be compared fairly, in context. Evaluation also includes the **assessment of program outcomes against established standards** and identifying which specific programs, institutions, or delivery systems are actually poor performers and which represent best practices.

Policy Recommendations

Finally, after the data are evaluated relative to established standards, **policy recommendations** must be formulated to improve those programs which are under-performing, to assist or sanction non-performers, and to promote "best practice" models which consistently exceed standards. It is not enough to realize that a poorly performing program needs to be changed nor should it be assumed automatically that "best practices" which worked for one subpopulation in a specific labor market will produce similar successes among other subpopulations or under different labor market conditions.

Effective and efficient policy responses cannot be devised unless the factors which determine program performance are understood. Once program performance has been rated on the basis of applicable standards and adjustments, it is vitally important to **explain** the direct effects of demographic and intervening factors on the outcomes that have been documented, the inter-relationships among those factors, and the interaction effects of the political and

Types of Measures

1. Inputs - Characteristics of subjects antecedent to or at the time they enter a program offered by a service provider; *inputs* also may be used to label the resources at the provider's disposal and constraints on the delivery of services—commonly factors which service providers can measure for themselves without requiring the assistance of an external follow-up entity. (Example: equity of access.)

2. Processes - The actual services, treatments, or interventions and how they were delivered—commonly factors which service providers can measure for themselves without requiring the assistance of an external follow-up entity. To some extent, process measures have, in the past, been taken as empirical indicators of the *quality* of services provided. (Example: student-teacher ratio.)

3. Outputs - Attributes or characteristics of subjects at the point when they exit a program or when services are terminated—commonly factors which service providers can measure for themselves without requiring the assistance of an external follow-up entity. (Example: graduation rate.)

4. Outcomes - What happened to subjects after services were provided—variously conceptualized as the *"impacts," "payoffs,"* or *"returns on the investment made in service delivery."* (Example: post-exit earnings.)

economic environment in which programs are operated. In short, based on hard performance data and rigorous analysis (rather than conjecture, anecdotal information, and marketing hyperbole), we must know which interventions are most likely to work for whom and under what conditions. Then and only then can policy-makers, administrators, and service providers determine how to tailor programs to customers' needs and how to maximize returns on the investment of taxpayers' dollars.

While data collection, performance evaluation, and policy response processes are codependent and inseparable in the broadest sense, each is sufficiently unique that they should not be performed by the same entity. The data collection process, for example, has specific and limited objectives. Data collection seeks to maximize technical proficiency by linking as many automated administrative and other databases together as possible to identify the full range of potential outcomes and participants for the least cost. The concerns of the data collection (or follow-up) entity include adequately defining the measures used so each variable is a meaningful representation or proxy for its measurement objective. The technical aspects of follow-up includelinking databases, preserving individual participant level record confidentiality and security, negotiating data sharing agreements, and expanding the range of coverage until all significant possibilities are exhausted.

Noticeably absent from this litany of concerns are value judgments, sanctions, and policy-setting activities. In fact, if the data collection entity is perceived in the role of making judgments or enforcing policy, its relationship with the various agencies and service providers on whom it depends for participant information and outcomes data are jeopardized. If the data collection entity is feared or mistrusted, partner agencies and service providers may attempt to thwart the negotiation of data exchange agreements. They may "sandbag" their data and "cook their numbers," in response to perverse incentives or otherwise second-guess the fiscal implications of subsequent program evaluations. They may drag their feet and fail to deliver program participant information in time to take advantage of very narrow windows of opportunity in the record linkage schedule. The follow-up entity responsible for gathering outcomes information, therefore, must remain independent and detached from program evaluation and policy-making (other than policies regarding data quality control, data sharing, and data security). In short, the function of the follow-up entity is to deliver valid, reliable, and timely information at a reasonable cost.

On the other hand, the program evaluation function is much more politically sensitive. Evaluation includes setting standards which invariably attaches labels of "exceeding standards," "outstanding," or "falling below standards." Such labels have implications in terms of public perception, funding, and program control. Programs which exceed standards often receive bonuses and incentive payments. Service providers operating under performance-based contracts may be subject to having some portion of their payments withheld or may forfeit status on a certified vendor list because of non-performance. Administrative entities may be subject to sanctions—even reorganization—if they consistently fail to meet performance standards. While program performance can be described in neutral

Types of Studies

1. Snapshot - A type of research design that gathers data about former participants at a single point in time; while adequate for several purposes, snap-shot studies cannot measure change over time (see Longitudinal Design).

2. Longitudinal Design - Research conducted on the same subjects at two or more points in order to assess changes in their behaviors, attitudes, experiences, or achievements over time; in employment and training follow-up, longitudinal designs are used to assess such things as learning gains, delayed or long-term program outcomes, earnings gains, and decreased welfare dependency.

3. Simulations - The process of identifying possible outcomes or policy proposals through "what-if?" analysis or testing of multiple hypothetical scenarios, e.g., if the state requires a placement rate of 70 percent, how many programs will likely fall above and below such a standard?

fashion, decisions about where to draw the lines for performance standards in the evaluation arena give rise to arguments based on self-interest. In the data collection phase, disputed claims and findings are settled through data dialog according to recognized statistical rules. In the evaluation phase, hard data may be considered less germane than well-reasoned estimates about who will be affected and best guesses about how stakeholders will react.

While program evaluation has implications which will affect the self-interest of agencies and service providers, many of the inherent activities can be done in a neutral and detached fashion. For example, in running a simulation to determine what happens if the standard for post-program employment is raised or lowered by 10 percent, the analyst can (and should) be "blind" to the parties affected. That is, codes or names identifying specific service providers should be stripped from the database during a simulation to determine how many (but not *which*) would fall into the "failed" and/or "outstanding" categories. Politicization of the evaluation process comes into play when the stakeholders are identified and consequences are about to befall them. It is at this point that value judgments must be made and potential repercussions must be weighed.

Devising policy responses, thus, is the most partisan of the three processes. Neither the technicians who gather data nor the analysts who concoct all the simulation scenarios are in a position to make policy because, ultimately, such decisions are determined according to the will of the people—to the extent that elected officials and their political advisors can decipher it.

The will of the people may change dramatically over time. What does not change, however, is the need for sound data to drive informed choice. It is for this reason that Texas has drawn a careful division of labor. Those who collect the data are bound by the widely recognized rules of sound research and descriptive statistics. Those who make policy must wrestle with the political ramifications of each decision. Those who do program evaluation have a foot in both arenas. On the one hand, they use their imagination or take their instruction from the policy-makers in determining what simulations are worth examining but, thereafter, follow the rules of statistical inference to take each scenario to its logical conclusion.

The division of labor, therefore, parallels the distinctions made in the philosophy of science.

- The follow-up entity that collects outcomes data is engaged in description: here is what happened.
- The program evaluation unit is engaged in prescription: if you want to achieve A, do X; if you want to achieve B, do Y.
- Policy-makers are engaged in making the final choice: is it in the public interest to achieve A or B?

The success of the policy-makers in promoting what they believe to be in the public interest depends upon the imagination and the predictive validity of the program evaluators' simulations. The soundness of the program evaluators' models and simulations, in turn, depends upon the validity, reliability, and timeliness of the information gathered by the data collection entity. Each is governed by a separate kind of logic and professional responsibility. The differences between these functions—albeit sometimes subtle—dictates that they should be performed by separate entities. Collectively, they lay the foundation for a comprehensive performance measurement and evaluation system.

THE TEXAS STATE OCCUPATIONAL INFORMATION COORDINATING COMMITTEE

TEXAS SOICC

3520 Executive Center Drive, Suite 205,
Austin, Texas 78731-1637
(512) 502-3750 • FAX (512) 502-3763
Toll-Free Career Information Hotline:
1-800-822-7526 in Texas
Internet: www.soicc.capnet.state.tx.us

TSOICC Member Agencies: Texas Education Agency, Texas Workforce Commission, Texas Department of Commerce, Texas Rehabilitation Commission, Texas Higher Education Coordinating Board.