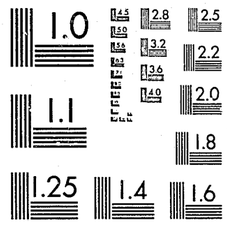


THE OSPREY COMPANY
2404 San Pedro Ave.
Tallahassee, FL 32304

PERFORMANCE MEASUREMENT THEORY FOR CORRECTIONS

Final Report
Grant 80-IJ-CX-0033

FINAL REPORT
Performance Measurement Theory for Corrections
Grant 80-IJ-CX-0033

This research focuses upon three issues related to measuring the performance of adult-offender related post-sentencing activities in probation and parole agencies:

1. What are the critical operations in probation/parole upon which performance measures ought to focus?
2. What measures can different constituent groups agree upon as being adequate measures of performance?
3. How does the relative importance of different performance dimensions vary among constituent groups and over time?

Our findings are briefly summarized below. Working papers that describe the research procedures and results in detail are identified in brackets following the summary of each issue. Eight of the working papers were written for publication in various journals. These articles are attached to this report, with notations as to which journal each article was submitted.

Operations upon Which Performance Measures Ought to Focus

We examined in depth the activities that officers at five probation/parole agencies performed when supervising probationers and parolees. Five basic operations common to all agencies were intake, case assignment, supervision, violations and terminations. Specific activities were classified as being enforcement, rehabilitation, or administration oriented. After examining the amount of time officers spent on fourteen critical activities related to these orientations, we found that four agencies most closely fit the control model of agency practice and the fifth more closely fit the passive model. [WP82-1]

Constituent Views of the Adequacy of Performance Measures

A national survey was conducted to determine the extent to which probation and parole administrators, criminal justice researchers, and oversight staff could agree upon specific performance measures for probation/parole agencies having different orientations. Agreement between oversight staff, administrators, and researchers was moderately high, with correlation coefficients on ratings of measurement importance ranging from .60 to .71, depending upon the probation/parole agency's orientation and the constituent groups whose ratings were compared.

Both administrators and oversight staff saw the greatest difference in measurement requirements being between the rehabilitation- and passive-oriented agencies. Oversight staff saw the enforcement-oriented agency requirements as being equally similar to both the passive- and rehabilitation-oriented agencies. Researchers and administrators, however, saw the enforcement agency's requirements as being more like the passive-oriented agency's than the rehabilitation-oriented agency's requirements.

Out of the sixty-five measures assessed, only four were judged appropriate for all three agency orientations by all three constituent groups. Three of these measured benefit and the fourth measured quality of service.

Promising statistical methods for weighting and aggregating performance measures to form a single indicator of overall performance include the performance ratio model, linear programming model, cost and production function models, and five multicriteria decision techniques. [WP82-7; WP82-9; WP83-3; WP83-5; WP83-6; WP83-7: WP83-8]

Relative Importance of Different Performance Dimensions

From a national sample of funders, researchers, and practitioners, we elicited preferences about the relative importance of six dimensions related to the performance of probation and parole agencies: quantity of output, quality of output, efficiency, equity, benefit, and cost-effectiveness. Regardless of the type group, they generally rated benefit and quality as being substantially more important than quantity and efficiency. These findings suggest that research priority should be given to developing benefit and quality measures.

The greatest variation in importance ratings occurred for the equity and cost-effectiveness dimensions of performance. Researchers assigned 22% of the total weight to equity, while funders assigned only 12% to equity. Funders, on the other hand, assigned 20% of the total weight to cost-effectiveness, while researchers assigned only 12% to that dimension. These differences may be large enough to have practical significance when using them to aggregate performance measurements on individual dimensions for purposes of ranking agencies or comparing their performance over time. [WP82-8; WP83-1; WP83-2; WP83-4]

Abstracts of Articles Submitted to Journals

1. Accountability for probation and parole agencies: Can administrators and oversight bodies agree on the terms?

   A national sample of probation and parole administrators and oversight staff rated performance measures that corresponded to several dimensions of accountability for program implementation. These dimensions addressed responsibility for both processes and policy outcomes. Overall, the level of agreement between the two groups was moderately high. A majority of both groups selected three benefit measures and one measure of service quality as important for assessing the performance of three different hypothetical agencies. The two groups did differ, however, on how important they believed some of the types of measures were for assessing agency performance. Administrators rated measures that serve to diagnose operations problems higher than did oversight staff. Oversight staff rated more service quality measures as important than did administrators.

2. Three perspectives on performance measurement: funders, practitioners, and researchers

   Agency performance may be described in terms of quantity and quality of output, equity, efficiency, benefits, and cost-effectiveness. A national survey indicates that practitioners, funders, and researchers differ in terms of how important they believe these different dimensions are in describing the performance of probation/parole agencies. This article applies their different perspectives about performance measurement to five probation/parole agency performance summaries. It then examines the resulting performance measurements and considers in what sense these differences in perspectives have practical significance.

3. Judging the performance of alternative corrections policies: a review of five techniques

   Priority-setting methods for policy boards must address both multiple criteria for choosing among alternative policies and the differing values of individual board members. Five techniques that might be appropriate for policy boards are decision analysis, simplified multiattribute rating technique, implicit multiattribute rating technique, analytic hierarchy process, and social judgment theory. This paper applies each of these techniques to three policies aimed at relieving prison overcrowding. Factors to consider before using one of these techniques include what roles board members and staff should play, how individuals' opinions should be aggregated, and whether the political conditions exist that make it feasible to use the technique.

4. Developing standards for interpreting agency performance: an exploration of three models

   Interpreting an agency's performance requires comparing its actual performance to a standard. The performance ratio, linear programming, and cost function models may each generate standards for comparing an organization's performance. This article shows how each model could be used to develop standards for probation and parole agencies. It then discusses the requirements for using each model and the model's advantages and disadvantages.

5. Efficiency in corrections agencies

   Most corrections agencies are in the formalization and control stage of their life cycles. The rational goal model is therefore the appropriate perspective from which to consider the performance of these corrections agencies. One of the important performance dimensions consistent with this model is efficiency. Production and cost functions may be useful techniques for developing a measure of overall efficiency for some types of corrections agencies. Cost functions were applied to time series data for twenty-one prisons and five probation and parole agencies. Using data on costs and output quantity and quality, the functions estimated the overall efficiency of each agency. Production and cost theory proved much more useful for analyzing the performance of large scale prisons than probation and parole agencies.

6. Performance measures for budget justifications: developing a selection strategy

   Performance information is important to the successful implementation of rationalistic budget reforms. Selecting "good" performance measures requires taking into account the context in which these measures will be used. A tool for systematically considering factors that render performance measures adequate or inadequate for a given situation is described. Corrections administrators can develop a measurement assessment strategy based upon this tool that is appropriate to their agencies' concerns and available resources.

7. Developing performance-dimension weights for assessing public-sector programs: method and contextual effects

   The sensitivity of weights to method and contextual factors was tested on a topic of interest to corrections policy-makers and administrators. Subjects used three methods - the analytic hierarchy process, social judgment theory, and the simplified multiattribute rating technique - to judge the relative importance of six performance dimensions for assessing public-sector programs. The dimensions judged were quantity and quality of program output, equity, efficiency, benefit, and cost. Subjects whose roles were funders, practitioners, researchers, or the general public judged the relative importance of these dimensions for assessing probation, medicaid, air pollution, and any/all programs. One measure of the sensitivity of weights elicited to the way the dimensions were defined was also tested. The dimension most sensitive to the factors tested was equity. In most instances the factor effects were not large enough to be statistically significant and interactions between factors were generally not statistically significant.

8. Integrating new methods for analyzing group decision making: social judgment theory, functional forms and random coefficient models

   Social judgment theory (SJT) is a method for eliciting opinions about the relative importance of multiple objectives or attributes. When SJT is used to elicit the opinions of individuals who form a group, one must consider by what method these individual opinions can be aggregated to represent the group's opinion. This paper suggests an appropriate functional form for analyzing opinions elicited by SJT and a method for combining individual opinions. It then applies the proposed model to the problem of establishing relative weights for six performance dimensions for a public sector agency. Analysis of data for individuals in two groups indicates that both interaction and quadratic terms are important in describing how individuals evaluate agency performance. Further, individual methods of agency evaluation are so diverse that a random coefficient model of valuation for the group as a whole is more appropriate than a fixed coefficient model.

Submitted by The Osprey Company

Gloria A. Grizzle

8/29/83

Gloria A. Grizzle, Project Director

Table of Contents

THE OSPREY COMPANY

ACCOUNTABILITY FOR PROBATION AND PAROLE AGENCIES:

CAN ADMINISTRATORS AND OVERSIGHT BODIES AGREE ON THE TERMS?

by

Gloria A. Grizzle

Working Paper 83-3
August 1983

submitted to Public Administration Quarterly

Prepared under grant 80-IJ-CX-0033 from the National Institute of Justice, U.S. Department of Justice. Views and opinions are those of the author and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

ACCOUNTABILITY FOR PROBATION AND PAROLE AGENCIES: CAN ADMINISTRATORS AND OVERSIGHT BODIES AGREE ON THE TERMS?

Accountability in public organizations connotes stewardship. In exchange for using public funds, the administrator is constrained to use them in a way that is responsive to the preferences of some person or group external to the administrator's own organization. To make accountability operative rather than simply symbolic, one must determine to whom the administrator accounts, for what s/he accounts, and by what instrument(s) responsiveness is assured.

Several instruments of accountability have been proposed. They include participative procedures, such as requiring consumer representation on policy boards or advisory committees (Etzioni, 1975) and giving notice and holding public hearings as a part of administrative rule-making (Smith, 1980-81). These procedures seek to ensure that actions the administrator will take in the future will be responsive to the public's wishes. Other procedures, such as detailed financial statements and program monitoring (Elling, 1983; Etzioni, 1975) require the administrator to account to oversight bodies for actions already taken.

Program monitoring is the accountability instrument considered in this paper. Some state governments are moving toward formalizing the monitoring process by requiring that agencies sign performance agreements or contracts with the legislature or governor in exchange for receiving appropriated

funds. The success of these performance agreements hinges upon whether agencies and their oversight bodies can agree upon meaningful measures and upon whether the oversight bodies subsequently reward agencies that meet the terms of their performance agreements or penalize agencies that do not.

The purpose of the research reported here is to learn whether administrators and their oversight bodies can agree upon specific measures by which programs should be monitored. The oversight bodies addressed are legislative and gubernatorial offices that review and make recommendations on agency budget requests. To make it possible to consider specific performance monitoring measures, yet be comprehensive in the types of measures considered, we focus upon a single agency type - probation and parole agencies. The scope of the research is limited to adult-offender related post-sentencing activities.

Drawing from the work of Etzioni (1975), McKinney and Howard, (1979) and Smith (1980-81), we define accountability as the administrator's responsibility to elected superiors for implementing policy in a way that conforms to the tasks, processes, and outcomes mandated by law. Types of outcomes considered important include equity in distributing those services, benefits, and exemptions to which individuals are entitled (Smith, 1980-81); efficiency in translating resources into products (Rosen, 1982; McKinney and Howard, 1979; and Smith, 1980-81); and effectiveness in translating products into desired outcomes (Thompson, 1980; McKinney and Howard, 1979).

Research Approach

Described below is the method used to obtain data about the

extent to which administrators and oversight bodies agree upon accountability measures for probation and parole agencies. The first section describes how potential measures were developed. Subsequent sections define the sampling frame and describe the survey instrument used to elicit opinions from probation/parole administrators and oversight staff.

The Performance Measures.--Considerable care went into developing a comprehensive set of performance measures tailored to probation and parole agencies. First, we reviewed thirteen models of criminal behavior from the disciplines of sociology, psychology, economics, and biology. These models ranged from radical theory in sociology to social learning theory in psychology to genetic/physiological theory in biology. Next we developed causal diagrams that related probation/parole agency activities to the intermediate, short-term, and long-term impacts expected, based upon these theories. Then we reviewed the corrections literature and identified several hundred potential measures that related to these causal diagrams.

To ensure that the measures would relate to actual goals and activities of the nation's probation and parole agencies, we made brief visits to eleven agencies scattered across the nation and discussed their programs, activities, goals and objectives with them. Then we made detailed observations of operations and conducted in-depth interviews with probation officers in five additional agencies, spending an average of seventeen days in each agency. From the resulting detailed information on agency operations we developed for each agency a set of performance measures specific to that agency's operations and sent them to

the agency administrator for review and comment.

Based upon the literature reviews and visits to probation and parole agencies, we next developed a set of measures that spanned the spectrum of goals, programs, and activities in the agencies we visited. To reduce this set to a manageable number for administrators and oversight staff to consider, we selected a subset of 65 measures that represented all the accountability dimensions encompassed by our definition of accountability.

Three types of measures were included to cover the task and process dimensions of accountability for policy implementation:

(a) quantity of output measures that describe the amount of an agency's direct products, i.e., the services rendered or the regulations enforced;

(b) quality of output measures that describe how well the agency is operating in terms of a variety of attributes, e.g., conformity to "good" practices, accuracy and timeliness of the work completed, the public's or offender's satisfaction with the service received;

(c) process diagnostic measures that help to identify what steps in the agency's processes or community linkages cause actual performance to differ from expected.

Four types of measures were included to cover the policy outcome dimension:

(a) equity measures that describe how fairly services are provided or regulations are enforced across different individuals or population groups;

(b) efficiency measures that estimate the cost per unit of output or overall agency efficiency compared to other agencies or

standards;

(c) benefit measures that describe the effect of the agency's actions upon the offender or others in society;

(d) cost-effectiveness measures that estimate the cost per unit of benefit.

On the survey instrument sent to the probation/parole administrators and oversight staff, measures were grouped under several categories. In the largest category were 37 measures that related to specific activities. Two to five measures were listed under each of twelve probation/parole activities. They were not labeled by type of measure (i.e.,as quantity, quality, process diagnostic, equity, efficiency, benefit, or cost-effectiveness). Each measure that could not be related to a single activity was listed under one of the following categories: agency cost and efficiency, agency characteristics, outcomes of agency activities.

Sampling Frame.--To assure that only people both knowledgable about probation/parole agencies and having a stake in measuring agency performance would be included in the survey, we focused upon two groups. First were administrators in probation/parole agencies. Two sources provided the sampling frame for these administrators - the 1981 edition of the Directory of Probation and Parole Agencies, published by the National Council on Crime and Delinquency and the 1977 edition of Expenditure and Employment Data for the Criminal Justice System, published by the U.S. Department of Justice and the U.S. Department of Commerce. We included in the sample all probation and parole offices with over 100 employees. We also drew a random sample of 100 offices

from those listed in the directory. Excluding duplications, 151 agencies were included in the administrators sample.

Second were the gubernatorial and legislative analysts responsible for reviewing and making recommendations on agency budget requests. Two other sources provided the sampling frame for this oversight group. The National Association of State Budget Officers membership list included the names of the executive budget officers for the fifty states. The 1981 edition of the Book of the States, Supplement #2, published by the Council of State Governments, listed the legislative budget offices for the fifty states. We drew a random sample of fifty offices from the total of 100. The survey instrument was directed to the executive or legislative analyst responsible for reviewing probation and parole agency budgets.

Survey Instrument.--Which measures are considered important may depend upon the goals and activities that an agency pursues. In order to assure that survey respondents had in mind the same agency characteristics when rating the performance measures, three profiles of hypothetical probation/parole agencies were developed and labeled agency A, B, and C. The purpose in developing the profiles was to present mixes of goals and activities actually encountered during the agency visits rather than to develop "pure" types as they might appear in the literature. Thus all agencies were portrayed as having multiple goals, with differences in goal emphases among the three agencies.

Agency A's goals were to assist the offenders in adjusting within the community, to provide treatment, and to protect the

community through reduction of recidivism. Its key activities were listed as being supportive counseling, effective probation and parole plans, psychiatric evaluations, job grooming, housing assistance, and monitoring offender progress. As a shorthand label we refer to this agency in the analysis below as the rehabilitation-oriented agency, although no such label was included on the survey instrument.

Agency B, referred to hereafter as the audit-oriented agency, had as its goals protecting the community; guiding and assisting the offender; and collecting fines, court costs, and restitution from offenders. Its activities included collection of payments, supervising offenders, keeping case books up to date, transfer paperwork, problem identification and referral, and apprehending offenders who violate the conditions of probation and parole.

Agency C, referred to as the enforcement-oriented agency, had as its goals protecting the community from the criminal activities of offenders in the agency's caseload and effecting rehabilitation through compliance with the conditions of probation/parole. Its activities included supervising offenders, problem identification and referral, investigation of possible violations of probation and parole conditions, and court appearances.

Each respondent judged the importance of each measure for each of the three hypothetical probation/parole ratings, making a total of 195 ratings per respondent. S/he rated each measure as being (a) not relevant for assessing the agency's performance, (b) relevant, but not important for assessing the agency's

performance, or (c) an important indicator of the agency's performance. The data analysis summarized below focuses upon the extent to which administrators and oversight staff agree upon which measures are important indicators.

Of the 151 instruments sent to administrators, 44 were returned. Seven of these were omitted from the analysis because respondents rated measures for only one of the three agency profiles. Two others were omitted because the instrument was filled in improperly, yielding 35 valid responses, or a 23% response rate. Of the 50 instruments sent to oversight staff, 26 were returned. Three were omitted because respondents rated measures for only one agency profile. Two more were omitted because the recipient sent the instrument to a probation/parole administrator and had him/her fill it out. The resulting response rate was 42%. With response rates at these levels, we do not assert that the survey results represent all probation/parole administrators and oversight staff. But we do believe that they reflect the opinions of the more interested and informed members of the populations sampled and therefore merit serious consideration. Indeed, we were surprised at the level of interest in performance measurement evidenced by comments and letters respondents attached to the instruments they filled out.

Findings

To compare the probation/parole administrators and oversight staff opinions, we first calculated the percentage of each group who rated each measure as being an important indicator of agency performance. Pearson product moment correlation coefficients were calculated for each agency profile to measure the degree of

association between the two groups' ratings. The coefficients for all three agency profiles are similar - .69 for both the rehabilitation- and audit-oriented agencies and .71 for the enforcement-oriented agency - and indicate a moderately high level of association between administrator and oversight staff ratings.

The scatter diagrams in Figures 1, 2, and 3 show the pattern of agreement for each agency profile. As these diagrams show, the measures do not break into well separated clusters, with one cluster being rated important by almost everybody in both groups and another cluster rated important by almost nobody in either group. Rather they form a continuum ranging from about 20 percent to 80 percent of both groups rating the measure as important. Measures falling below the diagonal are those which were rated important by a larger percentage of oversight staff than administrators. Measures above the line were rated important by a larger percentage of administrators than oversight staff.

What types of measures do both administrators and oversight staff believe are important for assessing agency performance? Because there is no clear clustering on the scatter diagram, we adopted an arbitrary cutoff of 60 percent to pursue this question. We identified for each profile those measures that at least 60 percent of both administrators and oversight staff rated important. Table 1 classifies these measures and compares their proportion to the distribution of all the measures included in

the survey. For this and the next table, we pooled the ratings for the three profiles. One would expect, if ratings had been made at random, that the proportions of measures selected that are of each type would be the same as their proportion of the total measures in the survey. The most noticeable difference is that half the measures that both groups agree are important are benefit measures, whereas only 25 percent of all the measures rated were benefit measures. The second highest proportion rated important were quality-of-output measures. No efficincy or cost-effectiveness measure was included in the 60 percent cutoff group.

Of the 65 measures rated, only four were rated important by at least 60 percent of both groups for all three agency profiles. Three of these measures fall into the benefit category: number and percentage of offenders who successfully complete their sentence (i.e, without violating their conditions of probation/parole), percentage of offenders arrested while on probation/parole, and percentage of offenders convicted of a new crime while on probation/parole. The fourth is a quality measure relating to general supervision of offenders: number of actual contacts per month per probationer compared to the prescribed number, broken down by level of supervision.

Continuing to use the 60 percent cutoff criterion, we find that the two groups rated ten additional measures as important for the rehabilitation-oriented agency. Five of these measures are benefit, three are quality, and one each is quantity and equity. Both groups also rated five additional measures as important for assessing the performance of the audit-oriented

agency. One measure was selected from each of the following categories: benefit, quality of output, equity, process diagnostic, and quantity of output. Of these additional measures, the process diagnostic and quantity ones were also rated important for the enforcement-oriented agency.

Next we look at those measures that at least 60 percent of administrators but fewer oversight staff rated important. Table 2 shows that the largest concentration of these measures falls into the process diagnostic category. The next largest concentrations are in the benefit and quality categories. It should not be surprising that diagnostic measures seem more important to administrators than to oversight staff. We would expect oversight staff to be more interested in summary measures of how well an agency is operating, whereas the administrator must concern himself with identifying and correcting specific operations problems. Examples of process diagnostic measures that more than 60 percent of administrators (but not oversight staff) rated as important include the number of offenders not referred to appropriate community resources for self-improvement or help because of inadequate community resources, the turnover rate of probation officers, the average time officers spend on selected activities compared to the agency's standard developed for each activity, and the ratio of units of work completed compared to the funded capacity.

Measures important to oversight staff but less so to administrators paint a different picture of agency performance. Quality measures dominate the set, followed by equity measures. Examples of quality measures that more than 60 percent of

oversight staff (but not administrators) rated as important include the percentage of agency effort devoted to offender supervision compared to the targeted percentage; the ratio of field contacts to office contacts; the average caseload size per officer compared to funded caseload size, broken down by supervision level. Examples of equity measures include the percentage of offenders counseled or treated who are rehabilitated, broked down by race, sex, and age group; the number and percentage of offenders violating conditions of probation/parole whose term is revoked, broken down by type of violation, age group, race, sex, type of offense, and length of term.

It is surprising that efficiency and cost-effectiveness measures were not rated as more important, especially by the oversight staff. The literature on accountability led us to expect that these dimensions of policy outcomes would be important to people. In addition, another study (Grizzle, 1983) that asked legislative and executive budget analysts to judge the relative importance of several performance dimensions instead of asking them to rate specific performance measures found that the budget analysts rated cost-effectiveness and quality about equally important. The cost-effectiveness measure rated lowest was rated important by 24 to 29 percent of the respondents, depending upon the agency profile for which the measure was selected. The cost-effectiveness measure that received the best rating (54 to 63 percent) was average supervision cost per offender successfully completing the probation/parole term.

To understand why the cost-effectiveness and efficiency

measures did not receive higher ratings, we analyzed the comments that respondents made about specific measures. These comments indicated that respondents considered the cost-effectiveness measures infeasible because they believed the data would be too difficult or impossible to obtain. The efficiency measure rated lowest (24 to 37 percent) was overall agency efficiency as a percentage of the most efficient, comparable agency in the state. Respondents indicated they believed it was impossible to obtain the data for the "comparable" agency.

Additionally, some administrators did not believe they should be held accountable for the criminal behavior of offenders. In the words of one administrator, "under no circumstances should agency performance be based on the continued criminal activity of offenders on the caseload. So often courts place people not appropriaate for probation on probation, many times against the probation officer's recommendation." This comment emphasizes the dilemma of wanting to hold administrators accountable for outcomes in situations where they do not have control over the factors needed to produce desired outcomes. As McKinney and Howard point out, the last two syllables in "accountable" form the word "able." Accountability should be accompanied by the authority to take the actions necessary to achieve the outcomes for which one is responsible.

Summary and Conclusion

A national sample of probation and parole administrators and oversight staff rated performance measures that corresponded to several dimensions of accountability for program implementation. These dimensions addressed responsibility for both processes and

policy outcomes. Overall, the level of agreement between the two groups was moderately high. A majority of both groups selected three benefit measures and one measure of service quality as important for assessing the performance of three different hypothetical agencies. The two groups did differ, however, on how important they believed some of the types of measures were for assessing agency performance. Administrators rated measures that serve to diagnose operations problems higher than did the oversight staff. Oversight staff rated more service quality measures as important than did administrators.

The emphasis given to benefit and service quality measures is consistent with a previous survey that asked respondents to rate performance dimensions rather than specific measures. Cost-effectiveness and efficiency measures, however, were not rated as high as previous research led us to expect. Respondents did not appear to reject the concepts of efficiency and cost-effectiveness as important dimensions of accountability, but some believed the specific measures rated were not feasible in terms of data collection requirements. Not surprisingly, some administrators did not believe they should be held accountable for outcomes that they could not control.

Based upon these survey results, we would expect that performance agreements between legislatures or governors and agencies would focus upon benefit and service quality measures. Oversight staff seem less interested in holding administrators responsible for quantity of output or conformance to specific processes. Neither group showed much enthusiasm about the efficiency and cost-effectiveness measures included in the

survey. Perhaps the greatest need for future performance measurement research is developing efficiency and cost-effectiveness measures that are both feasible in terms of data collection requirements and credible in terms of matching the administrator's responsiblity with his/her authority.

Figure 1

Ratings of Performance Measures
for the Rehabilitation-Oriented Agency

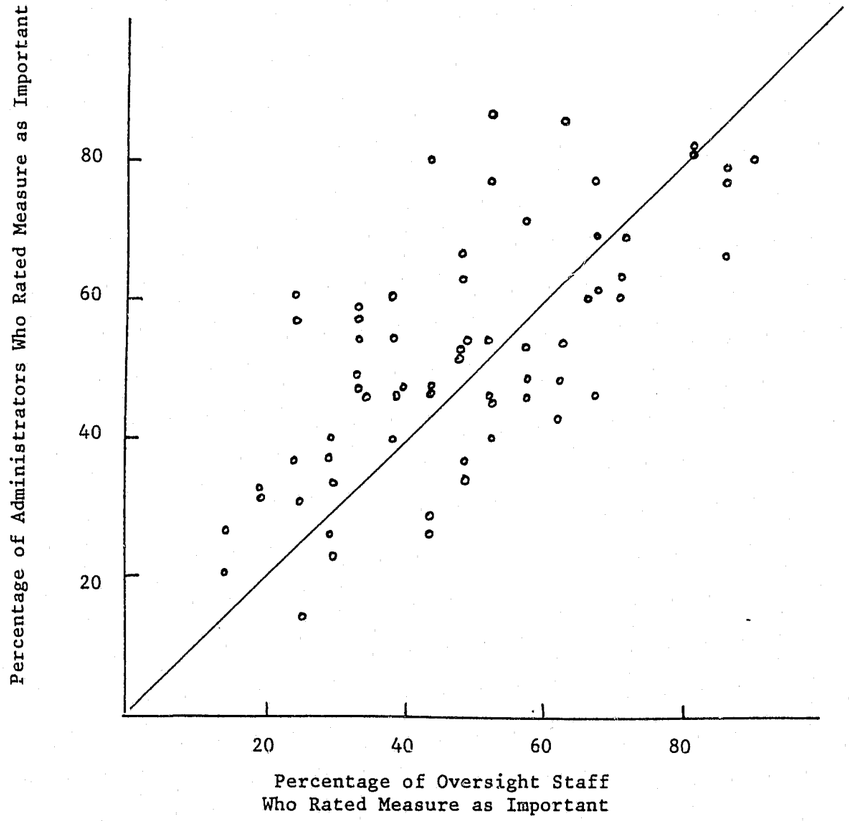


Figure 2

Ratings of Performance Measures

for the Audit-Oriented Agency

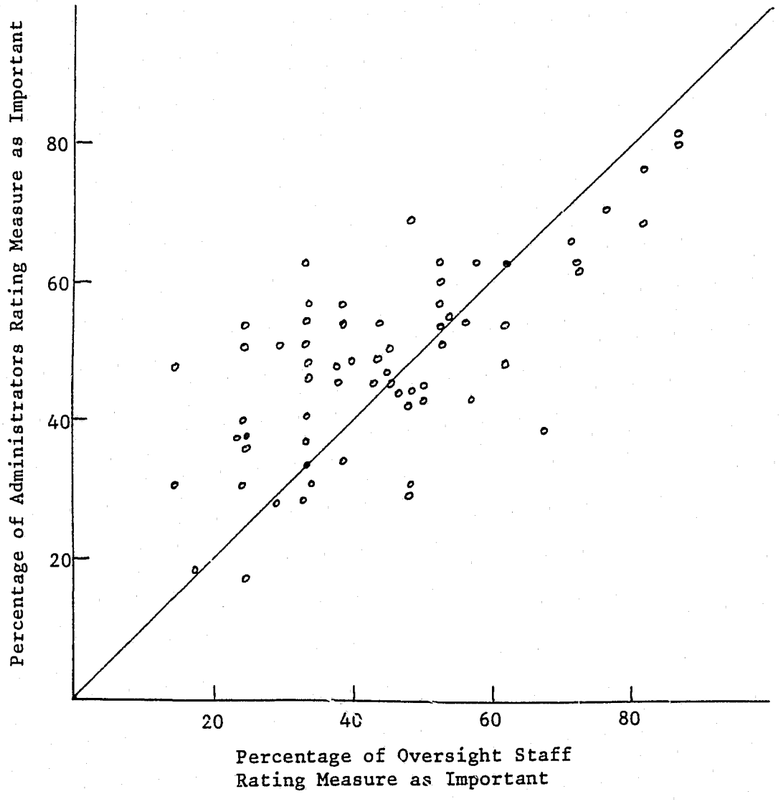


Figure 3

Ratings of Performance Measures

for the Enforcement-Oriented Agency

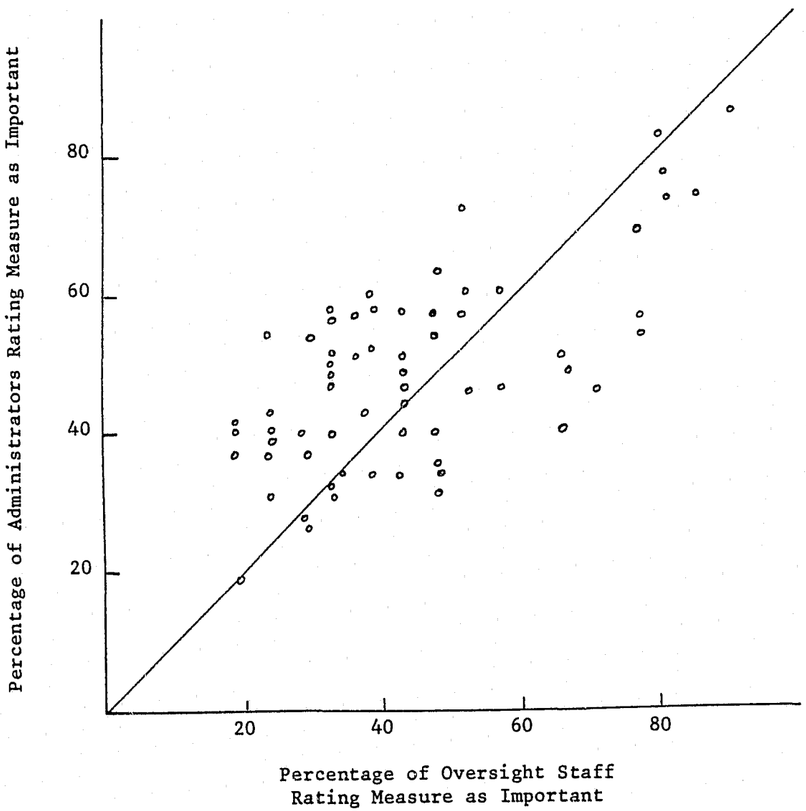


Table 1

Distribution of Performance Measures Rated as Being Important
by Both Administrators and Oversight Staff,
Compared to the Total Set of Measures Rated

Dimension	Total Set of Measures		Measures Rated Important	
	Number	Percentage	Number	Percentage
Quantity of output	27	14	3	11
Quality of output	33	17	7	25
Process diagnostic	36	18	2	7
Efficiency	12	6	0	0
Equity	24	12	2	7
Benefit	48	25	14	50
Cost-effectiveness	15	8	0	0
Total	195	100	28	100

Table 2

Distribution of Measures That at Least Sixty Percent of Only One Group
Rated as Being Important

Dimension	Administrators		Oversight Staff	
	Number	Percentage	Number	Percentage
Quantity of output	4	22	1	8
Quality of output	1	6	6	46
Process diagnostic	6	33	0	0
Efficiency	0	0	0	0
Equity	1	6	3	23
Benefit	4	22	1	8
Cost-effectiveness	2	11	2	16
Total	18	100	13	100

REFERENCES

Elling, Richard C. (1983) "Bureaucratic accountability; problems and paradoxes; panaceas and (occasionally) palliatives." Public Administration Review, 43:1, 82-89.

Etzioni, Amitai (1975) "Alternative conceptions of accountability; the example of health administration." Public Administration Review. 35:3, 279-286.

Grizzle, Gloria A. (1983) "Three perspectives on performance measurement: funders, practitioners, and researchers." Paper presented at the annual meeting of the American Society for Public Administration.

McKinney, Jerome B. and Howard, Lawrence C. (1979) Public Administration: Balancing Power and Accountability. Oak Park, Ill.: Moore.

Rosen, Bernard (1982) Holding Government Bureaucracies Accountable. New York: Praeger.

Smith, Brian C. (1980-81) "Control in British government: a problem of accountability." Policy Studies Journal, 9:14, 1163-1174.

Thompson, James Clay (1980) Rolling Thunder: Understanding Policy and Program Failure. Chapel Hill, N.C.: University of North Carolina Press.

THE OSPREY COMPANY

THREE PERSPECTIVES ON PERFORMANCE MEASUREMENT: FUNDERS, PRACTITIONERS, AND RESEARCHERS

by

Gloria A. Grizzle

Working Paper 83-2

May 1983

submitted to Public Administration Review

This research was funded in part by grant 80-IJ-CX-0033 from the National Institute of Justice, U.S. Department of Justice. Views and opinions are those of the author and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

THREE PERSPECTIVES ON PERFORMANCE MEASUREMENT: FUNDERS, PRACTITIONERS, AND RESEARCHERS

Agency performance can have many facets. Important aspects or dimensions of performance include both the quantity and quality of output, the equity with which these outputs are distributed, how efficiently these outputs are produced, what benefits result, and the cost-effectiveness of the resulting benefits.[1] This paper looks at the extent to which funders, practitioners, and researchers agree about the relative importance of these performance dimensions.

Why Learning the Relative Importance of Performance Dimensions Matters

Assessing overall agency performance by looking at individual performance measures can be difficult. Agencies may vary in terms of how well they perform on each dimension. Further, an agency's performance on each dimension may increase or decrease over time. Increasing performance on one dimension can sometimes be at the expense of decreasing performance on another dimension. For example, greater quantity of output may be achieved by lowering output quality. Or improvements in efficiency may be to the detriment of benefits to client groups. The picture becomes even more complicated when one attempts to compare the performance of different agencies over time.

To make it easier to compare performance over time or across agencies, then, one would like somehow to combine multiple performance measurements into a single indicator that summarizes

an agency's overall performance. Statistical models, such as the performance ratio and linear programming approaches,[2] provide methods for combining these measurements. However, these models require that someone determine the relative importance of the performance dimensions included in them. Thus, learning the relative importance of performance dimensions is an important step in measuring an agency's overall performance.

People may disagree about how important one performance dimension is compared to another. If people do in fact disagree, whose judgement about the relative importance of performance dimensions should be used when developing an overall measure of agency performance becomes an important question. It seems likely that a person's role might influence his/her perspective on performance measurement. For example, people who are responsible for allocating funds across programs might believe that cost-effectiveness is the most important dimension. People responsible for implementing programs, on the other hand, might believe that quality is more important than cost-effectiveness.

Research Method

To test this assumption, we elicited judgments about the relative importance of performance dimensions from individuals whose roles varied as follows:

> funders--executive and legislative staff who develop recommendations for the government's budget and appropriations.
>
> researchers--people at universities or other research organizations who study government agencies.
>
> practitioners--people who work at the administrative level in government agencies.

Several factors may affect the relative importance that people assign to different performance dimensions. We would hope that the most important determinant of assignments is the opinions that people actually hold. Other factors that may affect their assignment of relative importance include the method used to elicit their opinions and the way the task is presented to them.[3] To minimize the influence of these other factors, we described the task to all three groups of respondents the same way and used the same method for all three. We asked them all to think in terms of the same type of organization--a probation and parole agency.

The aim in selecting a sampling frame was to query people who were both familiar with probation/parole agency operations and who would be expected to have an interest in assessing probation/parole agency performance. We hoped that restricting the sample to such people would increase the diligence with which they completed the survey instrument and decrease the percentage of individuals polled who actually had no opinion about the relative importance of performance dimensions.

A separate sampling frame was developed for each of the three groups that comprised the national sample. The practitioner sampling frame consisted of administrators listed in the 1981 edition of the Directory of Probation and Parole Agencies, published by the National Council on Crime and Delinquency. The researcher sampling frame was constructed by selecting that subset of the American Society of Criminology membership list who gave an affiliation with a university or other research organization. We

drew a random sample of 100 people from each of these sampling frames.

Two other sources provided the sampling frame for funders. The National Association of State Budget Officers membership list included the names of the executive budget officers for the 50 states. The 1981 edition of the Book of the States, Supplement #2, published by the Council of State Governments, listed the legislative budget offices for the 50 states. We drew a random sample of 50 offices from the total of 100. We directed the survey instrument to the executive or legislative analyst responsible for reviewing probation/parole agency budgets.

The response rate for the three groups was as follows:

Funders--41 respondents, or 82% of the sample

Practitioners--43 respondents, or 43%

Researchers--48 respondents, or 48%.

Because of the small size of the funders sample, we sent out one follow-up letter to people who had not responded within one month to our original request. We did not follow up nonrespondents in the practitioner and researcher groups.

Figure 1 shows the survey instrument used to elicit judgments about the relative importance of performance dimensions. The respondent indicates his/her preferences through a series of pairwise comparisons. This format facilitates using Saaty's[4] analytic hierarchy process and corresponding statistics to analyze and interpret the survey results.

In a cover letter, respondents were told that the researcher was developing performance measures for probation/parole programs

and wanted to identify the types of measures that people thought were most important for judging the adequacy of agency performance. They were told that the survey findings would be used to set priorities on which types of performance measures to develop and test first. Finally, they were asked to judge the relative importance from their perspective as budget analysts, practitioners, or researchers.

Survey Findings

To analyze the survey data, each individual's response was first used to develop a vector of percentages that reflected his/her judgments about the relative importance of the six performance dimensions. The following method was used to produce this priority vector. Each individual's response was set up as a 6 X 6 matrix. If for each cell the performance dimension in the row was more important than the dimension in the column, the absolute value of the number checked by the respondent was inserted in the cell. If instead the dimension in the column was rated more important, the reciprocal of the absolute value was inserted in the cell. The lower lefthand half of the matrix is therefore the reciprocal of the upper righthand half of the matrix. (For illustrative purposes, one such matrix is reproduced in Table 1.) Next, the geometric mean was calculated for the six elements in each row, and the resulting priority vector was normalized so that the percentage would total 100%.[5]

To obtain group judgements about the relative importance of the six performance dimensions, we then calculated the arithmetic

mean for each of the six numbers in the individuals' priority vectors. The resulting vector for each group is shown below:

Dimension	Funders	Practitioners	Researchers
Quantity	8% (.75)	11% (1.00)	6% (1.00)
Quality	19 (.47)	22 (.41)	21 (.52)
Equity	12 (.75)	16 (.62)	22 (.59)
Efficiency	13 (.46)	11 (.45)	10 (.80)
Benefit	27 (.44)	28 (.43)	12 (.50)
Cost-effectiveness	20 (.50)	12 (.58)	12 (.75)

The number in parentheses is the coefficient of variability, obtained by dividing the mean into the standard deviation.

All three groups indicate that benefit is the most important dimension and quantity of output is the least important. The two major differences are the greater importance that funders place upon cost-effectiveness compared to practitioners and researchers and the greater emphasis that researchers place upon equity. These differences seem reasonable because cost-effectiveness is the decision criterion that proponents of economic rationality advocate for allocating resources across agencies or programs.[6] The researchers sampled come more from sociology and political science than economics and are therefore more concerned with who gets what services, a matter of equity, than with allocating resources on the basis of cost-effectiveness.

The coefficient of variability indicates the degree of homogeneity in individual judgments within each group. The smaller the coefficient, the greater is the consensus about the

dimension. Both funders and practitioners have the most consensus about the importance of the quality, efficiency, and benefit dimensions. Researchers show the most consensus about the quality, equity, and benefit dimensions. Quantity is the dimension for which there is the least consensus about its importance.

To compare the priority vectors obtained for the three groups, we used the root mean square deviation.[7] The equation for comparing two vectors that have six dimensions is

$$\sqrt{\frac{1}{6}\sum_{i=1}^{6}(a_i-b_i)^2}$$

where a_i is the percentage of the ith dimension in vector a and b_i is the percentage for the ith dimension in vector b. This root mean square deviation can range from 0 to 58. Zero represents identical vectors and 58 represents the maximum possible dissimilarity. Comparing the researchers and funders priority vectors, we find that the root mean square deviation is 5.5. The other pairs have slightly smaller root mean square deviations--3.2 for researchers compared to practitioners and 4.1 for funders compared to practitioners.

Some people may wonder whether the benefit dimension received the highest rating because it sounds good in the abstract. We tried to avoid such a response by grounding the performance dimensions in specific measures and including these measures on the form each respondent filled in. We also looked to see how benefit measures fared relative to other measures in a national

survey reported in another study.[8] In this other survey a majority in each of three similar constituent groups rated three of sixty-five measures as relevant and important for all three agency profiles. All three were benefit measures. The benefit measures were not labeled "benefit" but were grouped under "outcomes of agency activities."

As an additional check on the validity of benefit being judged as the most important performance dimension, we reviewed the legislative appropriations hearings for two states. These hearings were for the 1979 and 1981 Florida Senate and House subcommittees that dealt with corrections and the 1981 North Carolina House and Senate appropriations subcommittees that dealt with corrections. The approach was to transcribe each question that a legislator asked during these hearings and code each question as either relating or not relating to performance. Of the 127 questions that the legislators asked corrections agency staff about performance, 38% were questions about benefit. No other performance dimension contained as large a proportion of the performance questions.

Applying the Performance Weights to Performance Measurements

The priority vector for each of the three groups provides a set of weights that can be used to combine into a single performance measurement many measurements representing individual performance dimensions. In some respects the three constitutent groups have a similar pattern of performance dimension weights. All three judge benefit and quality as being more important than

efficiency and more than twice as important as quantity. Funders, however, judge cost-effectiveness to be more important than do the other two groups. Also, researchers judge equity to be more important than do the other two groups.

Are these differences large enough to have practical significance when using them? To better appreciate the effect that these differences might make when judging agency performance, we applied them to a set of performance measurements for each of five probation/parole agencies. The performance measures used to represent each performance dimension follow:

Quantity: Number of offenders supervised.

Quality: Percentage of referrals followed up.

Equity: Percentage of offender problems identified that resulted in referrals to obtain help.

Efficiency: Annual cost per offender supervised.

Benefit: Number of early and regular terminations as a percentage of total terminations.

Cost-effectiveness: Cost per successful termination.

The performance measurements for each agency were scaled so that the best possible performance would be scored 100% and the worst possible would be scored 0%. Where there was no external standard to define "best" performance, the agency which performed best for a given dimension was scored 100% and the other four agencies were scaled to that imputed standard. Table 2 shows these performance scores for the five agencies.

Table 3 shows three overall performance scores for each agency, using in turn the weights provided by the practitioners, the funders, and the researchers. Each agency's overall score was obtained by multiplying each performance measurement by the group's respective weight relating to that performance dimension and summing the resulting products. Although there is some variation in each agency's performance score, depending upon which group's weights are applied, their rank order does not change. Regardless of which group's preferences about the relative importance of the performance dimensions is used, agency A performs best, C second best, B third best, D fourth, and E worst. It is also worth noting that the same rank order would obtain if all performance dimensions were weighted equally.

As a second exercise, we developed performance measurements for hypothetical agencies. This exercise differed from the previous example in three ways. First, the number of agencies was increased to 80. Second, the measurements varied over a wider range than occurred for the five agencies whose performances are described above. Third, the performance weights applied were restricted to the two constituent groups that differed the most from each other--funders and researchers.

A random table provided the scores for each of 80 hypothetical agencies on each of the 6 performance dimensions. Each dimension's performance was allowed to range from 100, representing 10% of optimum performance, to 1000, representing 100% of optimum. To calculate an overall performance score for

each of the agency profiles, we multiplied each dimension's score times each group's weights and summed the products. For example, the overall performance score for agency profile #1 using the researchers weights is 620:

Dimension	Score		Weight		Weighted Score
Quantity	1000	X	8%	=	80
Quality	200	X	19	=	38
Equity	900	X	12	=	108
Efficiency	400	X	13	=	52
Benefit	600	X	27	=	162
Cost-eff.	900	X	20	=	180
Total score					620

Applying the funders weights to these same dimension scores gives a total performance score of 616 instead of 620.

Next we correlated the performance scores obtained by using the funders weights with those obtained by using the researchers weights. Figure 2 shows that there is a high, positive correlation between the two sets of performance scores. The Pearson product moment correlation coefficient is .93.

This high correlation again suggests that which group's weights are used might not make much practical difference. To pursue this possibility, we ranked the 80 agency profiles according to each set of performance scores and compared the two rankings. Table 4 shows that the differences in rank order range from 0 to 23. Profile #1, for example, would be ranked 17th out of 80, based upon either the funders or the researchers weights.

Profile #26, on the other hand, would be ranked first using the researchers weights but only 24th using the funders weights. The mean difference in rank order for the 80 profiles is 8.

Next we looked at the mean difference in performance scores. Total performance scores can range from 100 to 1000. The actual range for these 80 profiles is from 214 to 851 when using the researchers weights and 222 to 818 when using the funders weights. The mean difference between the two sets of scores is 38.

We conclude that a small change in the total performance score is enough to change the rank order. On the average a change of less than 4% (38 points out of 1000) is enough to change the rank order by 8 places. If such rankings were used to establish priorities among agencies for purposes such as program expansion or cutback, the choice of whose weights to use could materially affect the level of resources allocated to a given agency.

Summary and Conclusion

We elicited from a sample of funders, researchers, and practitioners their preferences about the relative importance of six dimensions related to the performance of probation/parole agencies. These dimensions were quantity of output, quality of output, efficiency, equity, benefit, and cost-effectiveness. On some dimensions judgments across dimensions resulted in similar rankings. Regardless of the type group, they generally rated benefit and quality as being substantially more important than quantity and efficiency. These findings suggest that research priority should be given to developing benefit and quality measures.

The greatest variation in importance ratings occurred for the equity and cost-effectiveness dimensions of performance. Researchers assigned 22% of the total weight to equity, while funders assigned only 12% to equity. Funders, on the other hand, assigned 20% of the total weight to cost-effectiveness, while researchers assigned only 12% to that dimension. These differences may be large enough to have practical significance when using them to aggregate performance scores on individual dimensions for purposes of ranking agencies or comparing their performance over time.

Whether they are large enough depends upon three factors:

(a) how much variation in performance actually occurs among the agencies being compared,

(b) how many agencies are compared, and

(c) how the overall performance measurements are used.

If the overall measurements are used to establish a rank order among agencies, then which group's weights were used would not matter for a few agencies if the variance in performance across agencies were similar to that for the five agencies sampled. The greater the number of agencies, however, the more likely a small difference in measurement will affect the rank order. If, however, actual variation were as great as that simulated for the eighty hypothetical agencies, then whether one used funder weights or researcher weights would make a difference in the rank order of the agencies.

Using the overall performance measurements instead of the resulting rank order of agencies could affect decisions even if the variation were no greater than that found in the five agencies described above. For example, the overall performance score for the poorest performing agency is 44 when weighted according to the funder preferences and 61 when weighted according to practitioner preferences. If these measurements were the basis for reimbursing an agency under a performance contract, whose weights were used could have a substantial effect on the size of payment the agency would receive. If one wants to use the overall performance measurement for such a purpose, who should establish the weights therefore becomes an important question that merits further study.

Another important problem needing study is how to elicit judgments when the judges have a personal stake in the outcome. In such situations, participants may be reluctant to reveal their true preferences. If an agency head knows, for example, that the performance weights he/she gives will be used to assess his/her agency's performance, he/she has an incentive to weight most heavily those dimensions that he/she believes the agency performs best on, regardless of which dimension he/she truly believes is most important. One way of countering this incentive is to have practitioners negotiate weights with their superiors, but this approach can also lead to gamesmanship in an adversary process. Each party may attempt to guess his/her opponent's weights and give extreme offsetting weights designed to bring the average weights into conformance with his/her own true weights.

The procedure used in this paper is adequate for making performance dimension weights explicit when people have no reason to conceal their preferences. The purpose of this exercise was to establish priorities for researching performance measures. Because this purpose is nonthreatening, these weights may be a truer indication of how these groups value different performance dimensions than would weights elicited as a part of an actual performance measurment process in a specific agency.

Footnotes

[1]Gloria A. Grizzle, "A Manager's Guide to the Meaning and Use of Performance Measurement," American Review of Public Administration, Vol. 15:1 (March, 1981), pp. 16-28.

[2]See Jin Eun Kim, "Cost-effectiveness/benefit Analysis of Postsecondary Occupational Programs: A Conceptual Framework," Planning and Changing, Vol. 11 (1980), pp. 150-165; Gloria A. Grizzle, "Using Statistical Models when Interpreting Probation Agency Performance: A Brief Exploration of Queueing Theory, Linear Programming, and Cost Function Applications," Presented at the annual meeting of the American Society for Public Administration, 1981.

[3]John C. Hershey, Howard C. Kunreuther, and Paul J. H. Schoemaker, "Sources of Bias in Assessment Procedures for Utility Functions," Management Science, Vol. 28:8 (August, 1982), pp. 936-954.

[4]Thomas L. Saaty, The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation (New York: McGraw-Hill, 1980).

[5]This approach is Saaty's method four for estimating the priority vector. Ibid, p. 19.

[6]See, for example, Verne B. Lewis, "Toward a Theory of Budgeting," Public Administration Review, Vol. 12:1 (Winter 1952), pp. 43-54.

[7]This statistic is recommended by Saaty op cit., p. 38.

[8]Gloria A. Grizzle and Karen S. Minerva, "Assessments of the Adequacy of Potential Performance Measures for Probation/Parole Agencies" (Tallahassee, Fla.: The Osprey company, 1982).

Figure 1

SURVEY INSTRUMENT

WHICH PERFORMANCE DIMENSIONS ARE MOST IMPORTANT?

Agency performance is a multidimensional concept. The term "performance" can include such dimensions as quantity and quality of output, equity, efficiency, benefit and cost-effectiveness. Opinions differ about the relative importance of these dimensions as indicators of agency performance. Definitions of each dimension and related performance measures are listed below.

Quantity of output refers to the amount of an agency's direct products, i.e., the services rendered or regulations enforced.
Examples: Number of contacts made with offenders
Number of investigations completed
Number of offenders referred to community resources

Quality of output refers to how well the agency is operating and encompasses a number of attributes, including conformity to "good" practices, accuracy and timeliness of the work completed, the public's or the client's satisfaction with the service received.
Examples: % of offenders who receive the level of supervision to which they were assigned
% of victims served by restitution program who are satisfied with the timeliness and adequacy of payment
Average elapsed time between need identification and referral of offender to a community resource

Equity refers to how fairly services or the enforcement of regulations are distributed among people. Common ways of breaking down services in order to compare their distribution among different groups of offenders include age, race, sex, extent of need, severity of offense or length of term.
Examples: % of offenders needing help who are referred to community resources, broken down by race, age group and sex of offender
% of offender problems identified for which help is obtained, broken down by whether obtaining help is a special condition of probation or parole
Average elapsed time between need identification and referral to a community resource, broken down by length of offender's term

Efficiency refers to the cost per unit of output.
Examples: Average cost per investigation completed
Average cost per office contact
Average cost per referral

Benefit refers to the effect of what the agency does upon the offender or others in society.
Examples: % and # of offenders who complete their term without violating a condition of probation or parole
% and # of offenders with drug or alcohol problems successfully rehabilitated
% and # of victims granted restitution who receive the full amount due them

Cost-effectiveness refers to cost per unit of benefit.
Examples: Average cost of securing employment for an offender
Average cost per alcoholic rehabilitated
Average cost for supervision of each offender who successfully completes a term without violation

THE OSPREY COMPANY
[illegible]

INSTRUCTIONS

Assume that your task is to determine the performance of a probation and/or parole agency. Use the matrix below to compare the importance of six performance dimensions as indicators of agency performance. Definitions of these dimensions appear on the lefthand side of this sheet.

Each row in this matrix compares two performance dimensions. For each row, check the column that most closely reflects your opinion of the importance of the performance dimension in the lefthand column compared with the performance dimension in the righthand column. For example, in the first row, a check in column +5 means that you believe quantity of output is strongly more important than quality of output. A check in column -3 means that quantity is moderately less important than quality. A check in column 1 means that the two performance dimensions are of equal importance as indicators of agency performance.

	Absolutely more important +9	Very Strongly more important +7	Strongly more important +5	Moderately more important +3	Equally important 1	Moderately less important -3	Strongly less important -5	Very Strongly less important -7	Absolutely less important -9	
Quantity	—	—	—	—	—	—	—	—	—	Quality
Quantity	—	—	—	—	—	—	—	—	—	Equity
Quantity	—	—	—	—	—	—	—	—	—	Efficiency
Quantity	—	—	—	—	—	—	—	—	—	Benefit
Quantity	—	—	—	—	—	—	—	—	—	Cost-effectiveness
Quality	—	—	—	—	—	—	—	—	—	Equity
Quality	—	—	—	—	—	—	—	—	—	Efficiency
Quality	—	—	—	—	—	—	—	—	—	Benefit
Quality	—	—	—	—	—	—	—	—	—	Cost-effectiveness
Equity	—	—	—	—	—	—	—	—	—	Efficiency
Equity	—	—	—	—	—	—	—	—	—	Benefit
Equity	—	—	—	—	—	—	—	—	—	Cost-effectiveness
Efficiency	—	—	—	—	—	—	—	—	—	Benefit
Efficiency	—	—	—	—	—	—	—	—	—	Cost-effectiveness
Benefit	—	—	—	—	—	—	—	—	—	Cost-effectiveness

Please check the category that most closely describes the position you hold:
___criminal justice practitioner ___researcher
___fiscal or budget analyst

Table 1

ILLUSTRATIVE MATRIX CONSTRUCTED FROM
AN INDIVIDUAL'S SURVEY RESPONSE

Performance Dimension	Quantity	Quality	Equity	Efficiency	Benefit	Cost-Effect.
Quantity	1	3	5	1/3	7	1/3
Quality	1/3	1	3	1/5	3	1/5
Equity	1/5	1/3	1	1/5	1	1/7
Efficiency	3	5	5	1	3	1
Benefit	1/7	1/3	1	1/3	1	1/7
Cost-Effect.	3	5	7	1	7	1

Table 2

Performance Measurements for Five Agencies

Performance Measure	Agency A	B	C	D	E
Number of offenders supervised	91%	72%	67%	100%	53%
Percentage of referrals followed up	89	84	57	74	99
Percentage of offender problems identified that resulted in referrals to obtain help	59	71	93	87	72
Annual cost per offender supervised	100	60	76	79	33
Number of early and regular terminations as a percentage of total terminations	90	88	96	54	55
Cost per successful termination	100	58	81	47	20

Table 3

Comparison of Overall Performance Scores for Five Agencies,
Applying Weights Elicited from Three Constituent Groups

Performance Dimension	Performance Measurement	Performance Score, Weighted by: Practitioners	Funders	Researchers
Agency A				
Quantity	91%	.10	.07	.05
Quality	89	.20	.17	.19
Equity	59	.09	.07	.13
Efficiency	100	.11	.13	.10
Benefit	90	.25	.24	.25
Cost-effectiveness	100	.12	.20	.12
Overall performance score		87%	88%	84%
Agency B				
Quantity	72%	.08	.06	.04
Quality	84	.18	.16	.18
Equity	71	.11	.09	.16
Efficiency	60	.07	.08	.06
Benefit	88	.25	.24	.25
Cost-effectiveness	58	.07	.12	.07
Overall performance score		76%	75%	76%
Agency C				
Quantity	67%	.07	.05	.04
Quality	57	.13	.11	.12
Equity	93	.15	.11	.20
Efficiency	76	.08	.10	.08
Benefit	96	.27	.26	.27
Cost-effectiveness	81	.10	.16	.10
Overall performance score		80%	79%	81%
Agency D				
Quantity	100%	.11	.08	.06
Quality	74	.16	.14	.16
Equity	87	.14	.10	.19
Efficiency	79	.09	.10	.08
Benefit	54	.15	.15	.15
Cost-effectiveness	47	.06	.09	.06
Overall performance score		71%	66%	70%
Agency E				
Quantity	53%	.06	.04	.03
Quality	99	.22	.19	.21
Equity	72	.12	.09	.16
Efficiency	33	.04	.04	.03
Benefit	55	.15	.04	.15
Cost-effectiveness	20	.02	.04	.02
Overall performance score		61%	44%	60%

FIGURE 2

CORRELATION OF PERFORMANCE SCORES, BASED ON RESEARCHERS VS. FUNDERS WEIGHTS, FOR HYPOTHETICAL AGENCY PERFORMANCE PROFILES

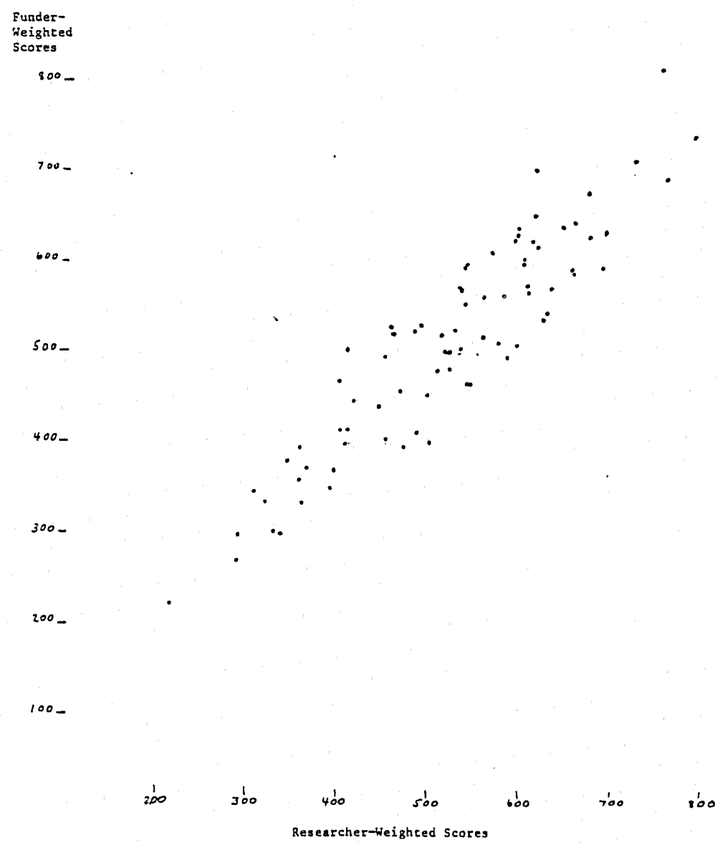


Table 4

DIFFERENCES IN RANK ORDER AND OVERALL PERFORMANCE SCORES OF 80 HYPOTHETICAL AGENCY PROFILES, USING RESEARCHERS WEIGHTS VS. FUNDERS WEIGHTS

Profile	Rank Order			Performance Score		
	Researcher-Weighted	Funder-Weighted	Difference	Researcher-Weighted	Funder-Weighted	Difference
1	17	17	0	616	620	-4
2	40	22.5	+17.5	533	597	-64
3	48	53	-5	506	481	+25
4	11	25	-14	655	590	+65
5	27	13	+14	592	633	-41
6	50	58	-8	493	453	+40
7	5	1	+4	750	818	-68
8	9	14	-5	672	630	+42
9	74	76	-2	335	304	+31
10	8	21	-13	685	598	+87
11	33.5	42	-8.5	556	519	+37
12	3	3	0	785	743	+42
13	45	46.5	-1.5	520	503	+17
14	32	18	+14	564	613	-49
15	52	63	-11	483	414	+69
16	33.5	32	+1.5	446	563	-7
17	4	6	-2	757	697	+60
18	22	26	-4	604	578	+26
19	66	54	+12	399	469	-70
20	77	73	+4	306	347	-41
21	25	44	-19	594	510	+84
22	49	48	+1	497	502	-5
23	70	68	+2	363	373	-10
24	31	43	-12	571	512	+59
25	53	39	+14	479	525	-46
26	1	24	-23	651	593	+58
27	29	15.5	+13.5	590	627	-37
28	7	12	-5	691	637	+54
29	28	50.5	-22.5	582	497	+85
30	72	70	+2	356	360	-4
31	6	4	+2	720	716	+4
32	14.5	34	-19.5	624	546	+78
33	47	41	+6	511	521	-10
34	23.5	20	+3.5	599	600	-1
35	14.5	35	-20.5	624	540	+84
36	68	69	-1	393	372	+21
37	79	78	+1	289	300	-11
38	64	65	-1	405	400	+5
39	76	74	+2	319	337	-18
40	20	15.5	+4.5	600	627	-18
41	16	9	+7	617	653	-36
42	78	79	-1	289	271	+18
43	44	52	-8	521	483	+38
44	51	36	+15	488	531	-43
45	73	67	+6	342	379	-37

Table 4 (continued)

Profile	Rank Order			Performance Score		
	Researcher-Weighted	Funder-Weighted	Difference	Researcher-Weighted	Funder-Weighted	Difference
46	54	66	-12	468	398	+70
47	35	55	-20	542	468	+74
48	10	7	+3	671	679	-8
49	55.5	37	+18.5	465	530	-65
50	69	72	-3	391	351	+40
51	63	46.5	+16.5	408	503	-95
52	75	77	-2	329	303	+26
53	57	40	+17	457	522	-65
54	26	11	+15	593	640	-47
55	36.5	29	+7.5	539	572	-33
56	2	2	0	829	775	+54
57	42	27.5	+14.5	530	573	-43
58	65	62	+3	401	415	-14
59	58	64	-6	449	405	+44
60	59	50.5	+8.5	447	497	-50
61	23.5	19	+4.5	599	606	-7
62	39	33	+6	534	555	-21
63	62	61	+1	408	418	-10
64	55.5	57	-1.5	465	459	+6
65	61	60	+1	414	449	-35
66	13	27.5	-14.5	631	573	+58
67	21	30	-9	607	569	+38
68	41	45	-4	532	505	+27
69	38	22.5	+15.5	535	597	-62
70	36.5	56	-19.5	539	467	+72
71	67	71	-4	397	355	+42
72	12	10	+2	641	641	0
73	30	31	-1	577	567	+10
74	46	49	-3	517	500	+17
75	71	75	-4	358	335	+23
76	80	80	0	214	222	-8
77	18	5	+13	612	704	-92
78	43	38	+5	524	526	-2
79	60	60	0	442	440	+2
80	19	8	+11	610	654	-44

THE OSPREY COMPANY

JUDGING THE PERFORMANCE OF ALTERNATIVE CORRECTIONS POLICIES: A REVIEW OF FIVE TECHNIQUES

by

Gloria A. Grizzle

Working Paper 83-7

June 1983

submitted to Administration and Society

Prepared under grant 80-IJ-CX-0033 from the National Institute of Justice, U.S. Department of Justice. Views and opinions are those of the author and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

JUDGING THE PERFORMANCE OF ALTERNATIVE CORRECTIONS POLICIES:

A REVIEW OF FIVE TECHNIQUES

This paper reviews several methods that policy makers might use for setting priorities among several corrections policies, based upon judgments about their overall performance. These methods apply to many general decision issues, such as developing comprehensive corrections programs, allocating funds among agencies, or evaluating the performance of agencies or programs. After describing each technique and comparing them, the paper explores the roles that policy board members and staff could play and how one could combine the preferences of individual board members whose values may conflict. Finally, it suggests the political conditions under which it would be practical for a policy board to use each method.

Introduction

Policy makers must frequently choose among alternatives that affect multiple objectives. Developing a community's capital improvement program provides one such example. Economic rationality might dictate that a single decision criterion (net present value or benefit-cost ratio) is sufficient for selecting projects. Local policy makers would probably feel it necessary to consider additional criteria, such as state and federal mandates, sources of funding, and the community's sense of urgency regarding different capital projects.

Multiple criteria require that the policy maker somehow

determine their relative importance in order to select the "best" projects. The literature on decision theory has devoted considerable attention to techniques that permit evaluating projects in terms of multiple attributes. Hwang and Yoon (1981) have recently reviewed 17 of these techniques. Most of these techniques have been developed for the use of individual decision makers. Such techniques may prove inadequate for plural policy-making bodies. These bodies must use a method that accommodates the different values of their individual members as well as multiple attributes of the alternative policies being considered.

Five techniques that might be suitable for plural policy making bodies are decision analysis, simplified multiattribute rating technique (SMART), implicit multiattribute rating technique (IMART), analytic hierarchy process (AHP), and social judgment theory (SJT). All these techniques divide the overall priority-setting process into several tasks. Generally, these tasks include identifying the attributes against which a policy should be assessed, determining effects by assessing each policy separately in terms of each attribute, determining weights by estimating the relative importance of each attribute, and combining the effects and weights so as to generate an overall rating for each project. Arraying these overall ratings in descending order sets the priorities for the policies analyzed.

Description of the Techniques

Decision Analysis

Decision analysis may be the best known of the priority-setting methods discussed here. Keeney and Raiffa's Decisions with Multiple Objectives: Preferences and Value Tradeoffs (1976) is the principal

explication of this technique. Public-sector appplications include designing police sectors for a city (Bodily, 1978), evaluating proposed sites for a pumped storage facility for generating electricity (Keeney, 1980), and selecting a site for a nuclear power plant for the Washington Public Power Supply System (Keeney and Nair, 1977).

Decision analysis is a compositional approach that requires the policy makers to trade off the importance of one attribute against another. It is the most complicated of the techniques to describe. Steps in arriving at an overall ranking of policy alternatives include the following:

1. Identify the policy alternatives that will be considered.
2. Identify the attributes of these policies that concern policy makers.
3. Determine the mathematical form that will be used to combine each policy's attributes into an overall ranking.
4. Determine the relative importance of the attributes.
5. Use the mathematical form, the attribute importance weights, and the policy's expected performance on each attribute to develop an equation that yields an overall score for each policy alternative.

Assume, for purposes of illustrating this technique, that a state government must alleviate overcrowding in its prisons. Three policies to achieve this end are being considered: build more prisons, parole more prisoners, and sentence more offenders to probation instead of prison. Assume further that three types of attributes concern the state's policy makers: cost to the state, restitution to victims, and prevention of future crimes through

deterrence. Cost may be broken down into two attributes. The first is direct cost to the state to implement each policy. For example, what would it cost the state to build and operate more prisons? Or to hire more probation officers to supervise more probationers? Second is the indirect cost for welfare assistance to families of prisoners and for taxes lost to the state on wages not earned while offenders are in prison. Deterrence may be broken down into three attributes. The first two attributes are specific to the offender. One is the prevention of future offenses that result in property loss to citizens. The other is prevention of future offenses that result in violence or physical harm to citizens. The third form of deterrence is general. Other people are deterred from crime by seeing the punishment meeted out to people who are convicted of commiting crimes.

For each of these six attributes a range of possible effects can be estimated. Direct cost might be stated in terms of millions of dollars a year to implement the policy. Let's assume that this cost could range from $10 million to $100 million. The worst level would be assigned a value of 0 and the best level would be assigned a value of 1. Future personal harm caused by the offender, on the other hand, might be stated in terms of the number of people a year who are injured by offenders while commiting future crimes. For this attribute, assume that the number might range from 200 to 1200 a year. The worst level, 1200, would be assigned a value of 0 and the best level of the attribute, 200, would be assigned a value of 1.

The third step in decision analysis is to determine the mathematical form that should be used to combine each policy's

expected effects in order to obtain an overall performance score. To keep the example simple, let us assume that one's preference for an attribute does not depend upon the levels at which the other attributes are fixed. For example, a policy maker would prefer a lower direct cost to a higher direct cost regardless of whether the number of people injured was at a high level or a low level. When this condition holds, the appropriate mathematical form is either additive or multiplicative. Let us assume that the appropriate form for this illustration is additive, meaning that for each policy alternative each attribute's weight can be multiplied by that attribute's value (utility) and the products can be summed to obtain an overall performance score for each policy alternative.

Determining the relative importance of the attributes, step 4, involves several tasks. We may begin by ranking the six attributes, accomplished by answering a series of questions. First, given that all six attributes are at their worst level, which attribute would you most like to have at its best level, assuming that the other five attributes remain at their worst levels? Assume that the answer is direct cost. Direct cost would then be ranked highest. Additional questions would be asked to determine the rank order of the other five dimensions. Assume that the rank order for the six attributes is as follows:

1. direct cost to implement policy
2. personal injuries caused by offenders
3. restitution to victims
4. property losses caused by offenders
5. indirect cost to the state
6. general deterrence

We next develop a utility function for direct cost that indicates the value that we place upon different levels of direct cost. Remember that the best case is valued at 1 and the worst case is valued at 0. This task now requires answering a series of questions. First, assume that there is a 50-50 chance of the cost being the worst case ($100 million) or the best case ($10 million). For what level of certain cost would you just as soon take the 50-50 lottery as the certain cost? Assume the answer is $55 million. The expected value of a 50-50 chance of either 1 or 0 is .5 ((1 + 0) / 2 = .5). Therefore we place in Figure 1 a dot on the graph that represents a value of .5 and a cost of $55 million. We can also place a dot at 1 value, $10 million cost to indicate the value we attach to the best case and another dot at 0 value and $200 million cost to indicate the value of the worst case.

By establishing intermediate ranges between these dots, we ask more questions to obtain additional points on the graph. For example, assume there is a 50-50 chance of the cost being $100 million or $55 million. For what level of certain cost would you just as soon take the 50-50 lottery as the certain cost? Suppose the answer is $77.5 million on the graph. By sketching in a line that best connects these five points we have a utility function that permits reading off the value we attach to any level of direct cost within the $10 million to $100 million range. This particular line happens to be straight, or linear, but frequently the utility function will be curved instead of linear.

The next task in determining the relative importance of the attributes is to establish tradeoffs between the most important attribute, direct cost, and each of the other five attributes.

Comparing the relative importance of personal injuries to direct cost requires answering the following question: Assume that there is a 50-50 chance of two events occurring. In the first event, direct cost would be $100 million and personal injuries would be at their best level, 200. In the second event, personal injuries would be at their worst level, 1200 and direct cost would be at level X. At what level would you set X so that you would be indifferent between the two events? Assume that the answer is $70 million. Now read the value for $70 million off the utility function in Figure 1. That value is about .3. If direct cost is valued at 1.0 at its best level, personal injury is therefore valued at 1.0 times .3, or .3, at its best level. A similar comparison with direct cost would be made to establish the relative weights for each of the other 4 attributes.

The final task in determining the weights is to translate the relative weights described above into absolute magnitudes. To do so we must answer another question: For what probability would you be indifferent between a policy costing $10 million and having the other 5 attributes at their least desirable level and an alternative policy consisting of a lottery yielding either all attributes at their most desirable level with the probability you chose or otherwise all attributes at their least desirable level? Assume the probability chosen is .6. The weight for direct costs should therefore be .6 and the other 5 attribute weights should be scaled relative to this weight. This scaling is accomplished by multiplying each of the relative weights by .6. For personal injury, the absolute weight would be .6 times .3, or .18. This task completes step 4 in the decision analysis technique.

The fifth step requires estimating how well each of the three policies would perform in terms of the six attributes. When we are uncertain about how well each policy will perform, probabilistic estimates can take this uncertainty into account. In this example, the greatest uncertainty surrounds the impact that each policy alternative would have upon general deterrence. We might estimate the probability that the number of people deterred from crime will be within each of several ranges on a scale. The scale could range from 0, meaning that no potential offender would be deterred, to 1.0 meaning that every potential offender would be deterred. For other attributes whose effects can be estimated with a high degree of certainty, a single point estimate for each policy alternative is sufficient.

Finally, assume that each policy's impact on each attribute has been estimated. A utility function must be developed in order to translate the policy's expected impact into a value ranging between 0 and 1.0. The procedure is the same as that already illustrated for direct cost in Figure 1. Table i illustrates such values for the three policy alternatives being considered. These values now need only be fit into a mathematical equation along with their respective weights to obtain an overall performance score for each policy alternative.

Recall that in step two we assumed that the appropriate form for this problem is additive. For each policy alternative, multiply its attribute effect shown in Table 1 times the respective weight for that attribute and sum the resulting products. For the policy alternative of building more prisons, the equation would be as follows:

Overall performance = .60x.3 + .18x.8 + .09x.1 + .07x.8 + .04x.2 +.02x.4 = .405

For the policy alternatives that would parole more prisoners and sentence more offenders to probation instead of prison, the overall performance values are .551 and .669, respectively.

Taking into account how well each policy would perform in terms of all 6 attributes, then, we would rank the policies as follows:

Policy	Overall Score	Rank
Sentence more offenders to probation	.669	first
Parole more prisoners	.551	second
Build more prisons	.405	third

Note that the heavy weight given direct cost causes the probation policy alternative to be ranked ahead of the prison alternative, which is the most expensive to implement. If property losses and personal injuries to citizens had been given more importance than direct cost to the state, then the overall score for the prisons alternative would have been greater than the overall score for the probation policy alternative.

Simplified Multiattribute Rating Technique (SMART)

In 1971 Ward Edwards proposed a rating technique designed to simplify the kind of judgments required by Keeney and Raiffa's decision analysis. He assumed that the organization, rather than a single individual, was the decision maker. Following this assumption, he partitioned the decision problem and looked to individuals with different expertise to render judgments for different parts of the problem. Applications included evaluating a community anti-crime program and evaluating the Office of Rentalsman as an alternative to the courts for handling landlord-tenant disputes (Edwards, 1980), ranking alternative desegregation plans

for Los Angeles schools (Edwards, 1979), and evaluating the street department's performance in Morgantown, West Virginia (Karako and Wolf, 1982).

For illustrative purposes we continue with the same three policy alternatives and six attributes in order to describe SMART. The first two steps, identify the policy alternatives and their relevant attributes, are the same for SMART as for decision analysis. SMART simplifies the third step - determining the mathematical form for combining the attributes. One simply assumes that the linear and additive form will be a good approximation to whatever the "true" form might be.

Determining the relative importance of the attributes first requires establishing the possible range of levels for each attribute. We can use the same ranges established for the decision analysis illustration. Direct costs range from $10 million to $100 million and personal injuries range from 200 to 1200. The worst point on each range is valued at 0 and the best point is valued at 100. Thus $10 million in direct cost would equal 100 and $100 million would equal 0. For personal injuries, 200 would equal 100 and 1200 would equal 0. Using SMART, one simply assumes that the change in the 0-100 value scale is proportionate to the change in the $10 million to $100 million cost and the 200 to 1200 personal injury scales. This assumption means that utility functions such as that shown in Figure 1 will always be a straight line and that one only needs to know the two end points on the range of the attribute to draw this line.

Given information about the range for each attribute, one is ready to decide how important each attribute is. There are several

ways of arriving at this decision. One three-step procedure is the following. First, arrange the attributes in rank order. Suppose that this rank order is the same as that used in the decision analysis example, where direct cost to implement the policy was deemed most important and general deterrence least important. Next, set the weight of the least important attribute equal to 10. Then compare each atttribute to the least important attribute in terms of how many times more important it is. Assume that direct cost is believed to be 10 times as important as general deterrence. The weight of direct cost is therefore 10 x 10, or 100. If indirect cost were 1.5 times as important as general deterrence, its weight would be 1.5 x 10, or 15. Third, normalize the weights so that they will total 100%. This step is accomplished by summing the six weights and dividing that total into each weight. This procedure is illustrated below:

Attribute	Weights	Normalized Weights
Direct cost	100/207 =	48%
Personal injuries	40/20 =	19
Restitution	22/207 =	11
Property losses	20/207 =	10
Indirect cost	15/207 =	7
General deterrence	10/207 =	5
Total	207	100%

The next step toward generating an overall performance score for each of the three policies is to estimate how well each policy will perform on each of the six attributes. This step is similar to that for decision analysis. Whereas in decision analysis the effects range from 0 to 1.0, in SMART the effects range from 0 to

100. We can use the same effects displayed in Table 1 simply by moving the decimal point two places to the right.

Finally, we multiply each policy alternative's effect for each attribute times that attribute's weight and sum the products to get the overall performance score. This step is the same as for the decision analysis illustration only because we assumed an additive mathematical form for the decision analysis example. In decision analysis, one first examines the decision maker's preference structure to determine whether the appropriate form is additive or multiplicative or some other form. In SMART, one always assumes the simplest form - a linear and additive one. For the policy alternative of building more prisons, the calculations are as follows:

Overall performance = 48%x30 + 19%x80 + 11%x10 + 10%x80 + 7%x20 + 5%x40 = 42.1

For the policy alternatives of paroling more prisoners and sentencing more offenders to probation, the overall scores are 51.3 and 62.4, respectively. The policies would then be ranked as follows:

Policy	Overall Score	Ranking
More probation sentences	62.4	first
Parole more prisoners	51.3	second
Build more prisons	42.1	third

Implicit Multiattribute Rating Technique (IMART)

Like SMART, the implicit multiattribute rating technique was developed specifically for plural policy-making bodies. As the name implies, attribute weights were not explicitly developed for the policy body as a whole. This method has been used to develop a

local drug abuse prevention and treatment program for Charlotte-Mecklenburg, North Carolina (Grizzle, 1973).

IMART, like decision analysis and SMART, requires identifying the policy alternatives for which priorities will be set and attributes in terms of which these alternatives will be compared. Unlike the two previous techniques, IMART permits the policy maker to ignore one or more of the attributes if he/she chooses. This provision permits the analysis to go forward even if all policy makers who will rate the policies cannot agree to a common set of attributes against which to assess the policy alternatives.

Once the alternatives and attributes have been identified, the effect that each policy alternative would have upon each attribute is estimated. These effects are then systematically displayed for the policy maker's review, perhaps in a format similar to that in Table 1. Each policy maker individually reviews the estimated effects but does not reveal to others the relative importance he/she attaches to them. Instead he/she makes an overall assessment of each policy alternative that permits establishing a rank order among them. The rank orders given to each policy are then arrayed, and the midpoint in the array is taken as the rank order for the policy making group. Suppose, for example, the policy body consists of five members. The rankings by the five members for the policy alternatives might be as follows:

Build more prisons - third, third, third, second, and second.

Parole more prisoners : third, third, second, second, first.

Sentence more offenders to probation - first, first, first, first, second.

Sentencing more offenders to probation would be given first

priority; paroling more prisoners, second priority; and building more prisons, third priority.

Analytic Hierarchy Process (AHP)

Thomas Saaty developed this technique in 1971 and has applied it to many policy issues since that time. These issues included setting priorities for transportation projects in the Sudan, analyzing alternative health care management policies in terms of their effect upon cost containment, setting land-usage priorities for different pieces of land, and setting resource priorities for a developing nation (Saaty, 1980).

As for the three techniques previously discussed, the first two steps in the analytic hierarchy process identify the policy alternatives to be considered and the relevant attributes. In step three the weights for the attributes are established by means of pairwise comparisons. Several tasks make up this step. First, a scale is constructed for indicating relative importance between attributes. This scale permits the policy maker to state the magnitude of difference in importance by a single-digit number. A typical scale follows:

Magnitude	Corresponding Definition
1	Two attributes are equally important.
3	One attribute is weakly more important than another.
5	One attribute is moderately more important than another.
7	One attribute is strongly more important than another.
9	One attribute is absolutely more important than another.

Unlike decision analysis and SMART, AHP does not require that the range of levels for each attribute be established before making

judgments about the attributes' relative importance. We compare each possible pair of attributes. Suppose our ratings are as follows:

Comparison	Magnitude
Direct costs and personal injuries	-3
Direct costs and restitution	-3
Direct costs and property losses	-9
Direct costs and indirect cost	-5
Direct costs and general deterrence	-7
Personal injuries and restitution	1
Personal injuries and property losses	-7
Personal injuries and indirect cost	-5
Personal injuries and general deterrence	-5
Restitution and property losses	-7
Restitution and indirect cost	-3
Restitution and general deterrence	-7
Property losses and indirect cost	+5
Property losses and general deterrence	+3
Indirect cost and general deterrence	-3

A plus sign to the left of the number means that the attribute on the left is more important than the attribute on the right. A minus sign means the attribute on the left is less important than the attribute on the right. According to the scale, then, the -3 opposite the first comparison means that the policy maker believes that personal injuries are weakly more important than direct cost.

These judgments about the relative importance of each attribute in each pair are next set into a matrix, as illustrated in Table 2. If the attribute in the row is more important than the attribute in

the column, the magnitude is expressed as a whole number. If the attribute in the row is less important than the attribute in the column, the magnitude is expressed as the reciprocal of the whole number. The numbers below the diagonal are reciprocals of the numbers above the diagonal.

Next, we summarize these numbers to arrive at a single number for each attribute that represents its weight. One method of doing so is to take the geometric mean of each row and then to normalize these means so that they sum to 1.0. Doing so for the matrix in Table 2 produces the following weights:

Direct costs	.03
Personal injuries	.05
Restitution	.06
Property losses	.46
Indirect cost	.14
General deterrence	.26
Total	1.00

To determine how well each policy alternative performs in terms of each attribute, we again use pairwise comparisons. The scale has the same magnitudes as before, but the corresponding definitions read as follows:

Magnitude	Corresponding Definition
1	Two policy alternatives have an equal effect.
3	One policy alternative is weakly better than the other.
5	One policy alternative is moderately better than the other.
7	One policy alternative is strongly better than the other.
9	One policy alternative is absolutely better than the other.

We now use this scale to generate six matrices. Each matrix summarizes our judgments about how well the three policy alternatives perform on a single attribute. Table 3 illustrates these matrices. These matrices are interpreted and summarized as explained in the paragraph describing how the attribute weights are determined.

The last step in the analytic hierarchy process is to combine the information about the policy alternative's effects on each attribute with the pertinent attribute weights and to summarize this information into a single overall performance score. We first create another matrix that consists of a column for each of the six attribute effect matrices. The columns consist of the normalized geometric means of the matrices in Table 3. Then we multiply this matrix by the attribute weights estimated earlier, as shown in Table 4. This calculation generates a single number for each policy that represents its overall impact ranked on a ratio scale. Because the prisons policy has the highest score, it would be ranked first. The scores for the parole and probation alternatives are almost identical.

Social Judgment Theory (SJT)

Decision analysis, simplified multiattribute rating technique, and the analytic hierarchy process are all compositional approaches that elicit attribute weights directly, either from direct scaling or paired comparisons. These weights are then multiplied by effects and summed to obtain an overall rating for each policy. Social judgment theory is a decompositional approach that infers the weights for each attribute from overall ratings.

Policy makers are first given a series of hypothetical policy alternatives and information that quantifies the extent to which each alternative affects each attribute. The policy makers make an overall rating for each hypothetical alternative. Attribute weights are then detected by analyzing these policy ratings. Applications include choosing the type of ammunition that police should use in their handguns (Hammond, 1976), selecting economic development policies for a county (Rohrbaugh and Wehr, 1978), planning a local government's budget (Steward and Gelbert, 1976), evaluating organizational performance (Rohrbaugh and Quinn, 1980), setting salary levels for individual faculty members at a state university (Roose and Doherty, 1978), setting priorities for an educational research institute (Adelman, Stewart, and Hammond, 1975), and setting a city's priorities on acquiring land parcels under the "Open Space" program (Hammond, Rohrbaugh, Mumpower, and Adelman, 1977, pp. 18-19).

As with the other four techniques, the first two steps in social judgment theory consist of identifying the policy alternatives and the attributes. The next step is to develop weights that indicate the relative importance of the attributes and, at the same time, a mathematical form for combining the attribute effects and weights into an overall score. This step involves three tasks.

First, one develops a set of hypothetical profiles that vary randomly in terms of how well a policy performed on each attribute. Figure 2 shows a sample of such profiles for the prison overcrowding problem. The bar opposite each attribute indicates how well that policy performed on that attribute. The worst possible performance

would be zero and the best possible 10. Second, the decision maker reviews each hypothetical profile and makes a judgment of overall performance by giving the profile a rating between 0 (worst) and 20 (best).

Third, a regression equation is fitted to these data. The dependent variable is the overall rating, and the attribute effects are the independent variables. As was the case for decision analysis, the relationship between changes in an attribute's effect level and changes in its value need not be linear. To keep this illustration simple, we assume that in this case the function forms for all the attributes are linear. Suppose that the regression equation that best relates the profiles to the judgment ratings is the following:

$$Y = .20X1 + .36X2 + .18X3 + .11X4 + .15X5 + .00X6,$$

where Y stands for the overall rating and the X's stand for direct costs, personal injuries, restitution to victims, property losses, indirect costs, and general deterrence, respectively. The coefficients are the weights that indicate the relative importance of the attriutes. Note that the coefficient for general deterrence is .00, meaning that the decision maker ignored this attribute when judging overall performance.

Armed with this regression equation, we can readily calculate an overall rating for any real policy, given estimates of its attribute effects. Assume that the effects for the three policy alternatives previously considered are the same as shown in Table 1, except that we move the decimal place one place to the right. For the alternatve of building more prisons, the overall rating would be calculated as follows:

$$Y = .20x3 + .36x8 + .18x1 + .11x8 + .15x2 + 0x4 = 4.84$$

For the other two policy alternatives, paroling more prisoners and sentencing more offenders to probation instead of prison, the overall scores would be 3.99 and 5.64, respectively. Given these attribute effects and weights, the three policies would be ranked as follows:

Policy Alternative	Overall Score	Rank
More probation sentences	5.64	first
Build more prisons	4.84	second
Parole more prisoners	3.99	third

Comparison of Techniques

Table 5 summarizes important characteristics that affect how useful these five techniques might be for policy boards. A technique is more likely to be used if it is simple, fits the policy maker's cognitive style, doesn't take much time, and does not demand information that is unavailable.

Decision analysis is the most complicated of the techniques described. It requires numerous judgments using the 50-50 lottery technique to establish the decision maker's preference structure and utility functions. SMART and AHP seem to make the least demands. The decision maker can focus upon one attribute at a time when evaluating policies and can establish attribute weights by comparing attribute pairs. SJT and IMART place a somewhat heavier burden on the decision maker by requiring that he/she take all attributes into

account simultaneously in order to judge each policy alternative's overall performance.

Policy makers may differ in terms of the cognitive style that is more comfortable for them. SJT and IMART require holistic judgments about a policy alternative's rating. The other three techniques are analytic, requiring a series of judgments about different aspects of a policy alternative.

Decision analysis and SJT require additional analysis to determine the form of the mathematical equation that will be used to aggregate weighted attribute effects. This step is not required in the other techniques, which should therefore take less time.

Finally, the analytic hierarchy process is the easiest to use when good information about the effects each policy alternative would have is unavailable. AHP allows comparisons to be made in terms of whether effects would be equal, better, or worse. The quantity of effect that is "equal," "better," or "worse" may be unknown. For decision analysis and SMART, ranges of effects must be established before tradeoffs are made between the attributes to establish their relative importance. SJT and IMART also presume that the estimates of effects can be obtained for each policy alternative.

Issues in Using These Techniques to Set Priorities

Three issues that need to be explored when policy boards use these techniques are (1) what role board members should play in analyzing policy alternatives, (2) how judgments of individual board members should be combined, and (3) whether it is politically acceptable to use any of these techniques.

Roles of Board Members

The role that the policy board plays is not inherent in the technique chosen. With any of these techniques the board may itself identify the policy alternatives and attributes to be considered, determine the attributes' relative importance, and determine the effect of each policy alternative on each attribute. Or, it may delegate any or all these tasks to staff or consultants. To clarify the board's options, we summarize below roles played in five public sector applications of these techniques.

1. Site selection for a power plant.--This application involved selecting a site for a nuclear power plant for the Washington Public Power Supply System (Keeney and Nair, 1977). The project team of consultants responsible for conducting the analysis identified the attributes in terms of which the alternative sites would be evaluated. Attributes included health and safety issues (radiation exposure, flooding, surface faulting), environmental effects (thermal pollution, sensitive or protected environments), tourism and recreation, and system cost and reliability. "Experts" from the project team determined the relative importance of the attributes. Based upon judgment and analysis of empirical data, the consultants also estimated each alternative's effect upon each attribute.

2. School desegregation.--This application ranked alternative desegregation plans for Los Angeles schools (Edwards, 1979). Staff and the consultant identified the outcome (or attributes) in terms of which the alternative desegregation plans would be judged. Examples of outcomes selected include the plan's effect upon racial-ethnic composition of schools, educational quality, community acceptance, and stability. They also determined the relative importance of these outcomes. School district staff then estimated how well each

desegregation plan would perform in terms of each outcome.

3. Drug abuse prevention and treatment.--In this application (Grizzle, 1973), a citizen committee and its staff jointly identified the attributes against which 44 projects were to be evaluated. The proponent of each of the individual projects that was considered for inclusion in the comprehensive program supplied his estimate of that project's cost, its target group and number of people whom the project would reach, the type of expected impacts on the target group, the percentage of total need met, and the likelihood that the anticipated impacts would be achieved. Each committee member then reviewed each project and its proponent's estimates and made his/her own estimates for each project. After reviewing the two sets of estimates of the projects' effects, each committee member ranked the 44 projects. In so doing, he was free to take into account whatever attributes he chose and to weight them however he chose in order to arrive at a holistic ranking.

4. Higher education.--This application considered the effects of seven higher education policies upon four objectives (Saaty, 1980, pp. 132-138). A group of 28 college-level teachers identified the objectives to which they believed higher education policies should contribute: prosperity, civil order, profit for industry, and perpetuation and power for industry. The relative importance of these objectives was established through a two-step procedure. The teachers first reached a consensus on the relative importance between each pair of objectives. Second, the teachers determined, also by consensus, how well each of seven policies would attain each of the four objectives.

5. "Open Space" land acquisition.--This application set a

city's priorities on acquiring land parcels under the "Open Space program (Hammond, Rohrbaugh, Mumpower, and Adelman, 1977, pp. 18-19). A local board in Boulder, Colorado, identified the attributes against which specific land parcels would be evaluated. These attributes included aesthetics, cost, location, availability, need for action, use potential, and contribution to protection of the environment. The consultants described forty hypothetical parcels of land in terms of their effects on these attributes. Each board member then scored each hypothetical parcel in terms of the overall desirability of acquiring it. The consultants then regressed these overall scores against the hypothetical parcels' effects on the attributes to infer the weights for each attribute. The mean of individual board members' weights were discussed by the board and modified by consensus. Once the attribute weights were established, the board members also individually rated each of the actual parcels of land that were to be acquired in terms of its effect on the seven attributes.

Table 6 summarizes the roles played in these five applications in terms of who identifies the attributes, who determines the attributes' relative importance, and who determines the effect of each policy alternative on each attribute. In the site selection application, the team of consultants does all these tasks, leaving to the policy board only the task of officially setting priorities based upon the consultant's recommendation. At the other extreme are the higher education and "Open Space" applications. In both these instances, the policy boards do all three tasks. In the school desegregation application, the board reviews and expands attributes and determines their relative importance but leaves the

determination of effects to its staff. In the drug abuse application, the board and staff jointly identify the attributes, the board members implicitly set their own weights individually when ranking projects, and the board members individually revise proponent's estimates of effects.

Consultants working with the policy board in the "Open Space application concluded that it was a mistake for the policy board to determine the effects and that this task could have been done better by technical experts (Hammond, Rohrbaugh, Mumpower, Adelman, 1977, p. 19). In a subsequent application they followed the logic espoused by Ward Edwards and had technical experts determine the effects for each policy alternative (Edwards, 1980). Where cause-effect theory and empirical data exist that permit estimating effects, it seems reasonable to have those effects estimated by whoever has the best information available, funds permitting. Where empirical data are absent or are of poor quality, one wonders who the expert is. In such situations, board estimates may be as good or better than anyone else's.

Combining Board Member's Judgments

If individual board members have difference opinions about the relative importance of attributes or the effects that each policy alternative would have, how can a single set of priorities for the board as a whole be developed? Three basic approaches may be taken to aggregating board members' judgments. The first, and most time consuming method, is to have the board sit as a group and through discussion reach a consensus. Consensus could be required about which attributes would be considered, their relative importance, and what attribute effects would result from each policy alternative if

implemented.

At the other extreme, each individual board members's judgment could be used to establish a separate overall performance score for each policy alternative. These scores could then be averaged to represent the collective opinion of the board. IMART prescribes this approach, except that the midpoint rather than the mean is taken as the board's ranking for each policy alternative.

The third approach is to obtain individual member attribute weights and effects. These weights and effects are then averaged and a single overall performance score calculated for each policy alternative. Saaty (1977) believes this approach is reasonable when individual judgments reflect indecision or possession of little information.

Political Acceptability of Using the Techniques to Set Priorities

Grizzle (1982) has noted political constraints that limit the degree to which policy boards can successfully follow the pure rational approach to decision making. In a political arena, managing information can be an important tactic in one's strategy to build a winning coalition. Making one's weights and estimated effects explicit may hinder rather than facilitate the coalition-building process.

It is instructive to note the outcome of the Los Angeles school desegregation policy analysis. Five of the seven board members provided weights, given a guarantee that individual members' weights would be kept confidential and only the average weights would be reported. These weights were used in analyzing eight alternative policies. The board adopted none of these policies. When the decision deadline arrived, the board adopted a compromise policy put

together shortly before the deadline. SMART was not used to evaluate this policy (Edwards, 1979, p.48). As another example, we quote J. G. Roche, reporting his application of decision analysis to the budget allocation problem of a small school district: "Under normal conditions, I don't believe it would be reasonable to expect that policy makers would allow their own preference structures to be communicated" (reported in Keeney and Raiffa, 1976, p. 376).

Under what political conditions are board members likely to make known their judgments about the relative importance of attributes? Making one's values explicit seems more likely when these conditions hold:

1. The constituencies that the board members represent share values.

2. The board's environment is friendly rather than hostile.

3. Publicizing value tradeoffs will not damage future support of the board or its policies.

4. Disclosing value tradeoffs will not exacerbate conflict and make agreement on any policy more difficult.

When board members share values, aggregation by either consensus or averaging members' weights would probably produce similar weights. When board members do not share values, reaching a consensus may prove impossible. Once they learn the rules of the averaging game, board members who do not share values may not give their individual weights in good faith. They may give extreme weights in order to offset their opponents' weights.

The ability to skew a policy board's average weights by individual board members giving deliberately extreme weights would probably be less of a problem with AHP and IMART. AHP limits the

range of possible weights from 1 to 9. IMART uses the median as the average, making the average less susceptible to being skewed by extreme weights than would be the case if the mean were used as the average. SMART, decision analysis, and SJT, on the other hand, place no limits on the range of weights that an individual member may assign to the attributes.

In the absence of shared values, it may be necessary to appeal to an "expert" to set the weights or to keep the weights implicit through such a technique as IMART. Either course of action may produce board decisions that are not optimal.

If board members do not participate in setting the weights, they will probably pay less attention to the priority rankings that the method produces when they bargain to obtain agreement on a policy that will satisfy a majority of the members. Using IMART, each member conceals information about his value tradeoffs, making it difficult for his opponents to counteract the effect of his priority rankings by making their rankings at the opposite extreme. It is unclear to what extent IMART's systematic process for considering policy attributes and effects before individuals rank policies will produce rankings that are different from simple bargaining.

Summary and Conclusion

Five techniques that policy boards may use to set priorities among alternative policies are decision analysis, simplified multiattribute rating technique, implicit multiattribute rating technique, analytic hierarchy process, and social judgment theory. Though these techniques differ in approach and implementation, they have several tasks in common. They identify the policy alternatives

to be assessed, establish a set of attributes that will be used to systematically assess each policy alternative, determine the relative importance of the attributes, determine the effect that each policy has upon each attribute, and combine this information into an overall performance score for each policy alternative.

All these methods seem technically adequate for policy boards to use when setting priorities. All are flexible in the division of tasks between the board and its staff or consultants. They do differ in terms of complexity, cognitive style, time required, and information demands - all factors that affect the likelihood that policy boards will use these techniques.

Finally, it must be remembered that techniques designed as neutral tools may not be used that way in a political environment. In some situations individual board members may subvert the intent of these techniques by manipulating their rankings to drive the board's priority rankings toward their own predetermined rankings. If individual policy makers have already decided which policy alternatives they prefer, then it may be a waste of time to use any of these techniques.

Table 1

Estimated Effects of Policy Alternatives

Attribute	Weight of Each Attribute	Build More Prisons	Parole More Prisoners	Sentence More to Probation
Direct cost	.60	.3	.7	.8
Personal injuries	.18	.8	.3	.4
Restitution	.09	.1	.2	.7
Property losses	.07	.8	.5	.4
Indirect cost	.04	.2	.4	.6
General deterrence	.02	.4	.4	.1

Table 2

Summary of Pairwise Comparisons of Attributes

Attribute	Direct Costs	Personal Injuries	Restitution	Property Losses	Indirect Cost	General Deterrence
Direct cost	1	1/3	1/3	1/9	1/5	1/7
Personal injuries	3	1	1	1/7	1/5	1/5
Restitution	3	1	1	1/7	1/3	1/7
Property losses	9	7	7	1	5	3
Indirect cost	5	5	3	1/5	1	1/3
General deterrence	7	5	7	1/3	3	1

Table 3

Matrices Showing Relative Effects Policy Alternatives Have upon Attributes

Direct costs

	Prisons	Parole	Probation
Prisons	1	1/5	1/9
Parole	5	1	1/3
Probation	9	3	1

Personal injuries

Prisons	1	5	9
Parole	1/5	1	3
Probation	1/9	1/3	1

Restitution

Prisons	1	1/3	1/9
Parole	9	1	1/3
Probation	3	3	1

Property losses

Prisons	1	7	9
Parole	1/7	1	3
Probation	1/9	1/3	1

Indirect cost

Prisons	1	1/3	1/7
Parole	3	1	1/5
Probation	7	5	1

General deterrence

Prisons	1	3	5
Parole	1/3	1	5
Probation	1/5	1/5	1

Table 4

Calculation of Overall Performance Score, Analytic Hierarchy Process

Attribute Weights			Attribute Effects							
				Direct costs	Personal injuries	Restitution	Property losses	Indirect cost	General Deterrence	
Direct costs	.03									
Personal injuries	.05									
Restitution	.06		Prisons	.28	3.56	.33	3.98	.36	2.47	
Property loss	.46	X	Parole	1.19	.84	1.44	.75	.84	1.19	=
Indirect cost	.14		Probation	3.00	.33	2.08	.33	3.27	.34	
General deterrence	.26									

Policy Alternative	Overall Score
Prisons	2.73
Parole	.94
Probation	.93

Table 5

Characteristics of Priority-Setting Techniques That Affect Their Usefulness for Policy Boards

Complexity

Most Complex	Intermediate	Simplest
DA	SJT	SMART
	IMART	AHP

Type Judgments Required

Analytic	Holistic
DA	SJT
AHP	IMART
SMART	

Time Required to Fit Mathematical Function

Most	Least
DA	SMART
SJT	IMART
	AHP

Applicability When Effects Cannot Be Quantified

High	Low
AHP	DA
	SJT
	SMART
	IMART

Table 6

Roles Played in Several Public-Sector Applications

Policy Application	Who Identified Attributes	Who Set Attribute Weights	Who Determined Attribute Effects
Site selection for power plant	Team of consultants	Experts from among consultants	Consultants
School desegretation	Consultant and staff, expanded by school board members	Individual school board members	School district staff
Drug abuse prevention and treatment	Citizen committee and staff	Individual committee members	Each project's proponent, revised by individual committee members
Higher education	College-level teachers	Teachers by consensus	Teachers by consensus
"Open Space" land acquisition	Board of trustees	Individual board members	Individual board members

Figure 1

Utility Function for Direct Cost

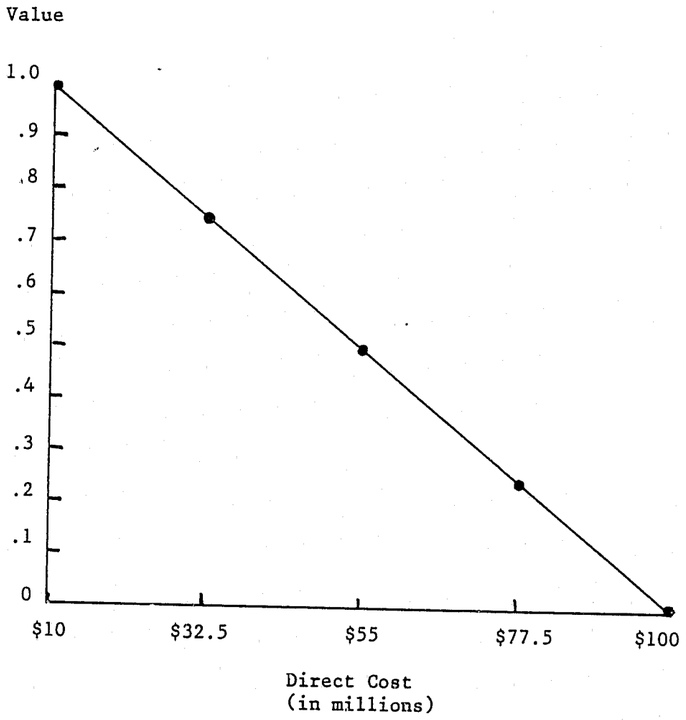


Figure 2

Profiles of Hypothetical Policy Alternatives

Alternative 1	Direct cost	XXXXXXXXXXXXXXXXXXX
	Personal injuries	XXXX
	Restitution	XXXXXXXXXXXXXXXX
	Property losses	XXXXXXXX
	Indirect cost	XXXXXXXXXXXX
	General deterrence	XXXXXXXXXXXXXXXXXX
		1 2 3 4 5 6 7 8 9 10
Alternative 2	Direct cost	XXXXXXXXXXXXXXXX
	Personal injuries	XXXXXXXXXXX
	Restitution	XX
	Property losses	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
	Indirect cost	XXXXXXXXXXXXXXXXX
	General deterrence	XXXXXXXX
		1 2 3 4 5 6 7 8 9 10
Alternative 3	Direct cost	XXXX
	Personal injuries	XXXXXXXXXXXXXXXX
	Restitution	XXXXXXXXXXXXXXXXXXXXXXX
	Property losses	XXXX
	Indirect cost	XXXXXX
	General deterrence	XXXXXXXXXXXXXXXXXXX

REFERENCES

ADELMAN, L.; STEWART, T. R. and HAMMOND, K. R. (1975) "A case history of the application of social judgment theory to policy formation." Policy Sciences, 6:137-159.

BODILY, S. (1978). "Merging the preferences of interest groups of efficiency and equity of service in the design of police sectors," in R. C. Larson (eds.). Police Deployment: New Tools for Planners. Lexington, Mass.: Lexington Books, 103-122.

EDWARDS, W. (1980). "Multiattribute utility for evaluation: structures, uses, and problems," in M. W. Klein and K. S. Teilmann (eds.). Handbook of Criminal Justice Evaluation. Beverly Hills: Sage, 177-215.

EDWARDS, W. (1979). "Multiattribute utility measurement: evaluating desegregation plans in a highly political context," in R. Perloff (ed.). Evaluator Interventions: Pros and Cons. Beverly Hills: Sage, 13-54.

GRIZZLE, G. A. (1982). "Plural policy-making bodies: decision strategies." Administration and Society, 14:81-99.

GRIZZLE, G. A. (1973). "Generating information for policy making in the field of drug abuse." Ph.D. dissertation. Chapel Hill, N.C.: University of North Carolina.

HAMMOND, K. R. (1976). "Externalizing the parameters of quasirational thought," in M. Zeleny (ed.). Multiple Criteria Decision Making, Kyoto 1975. Berlin: Springer-Verlag, 75-95.

HAMMOND, K. R.; ROHRBAUGH, J.; MUMPOWER, J.; ADELMAN, L. (1977). "Social judgment theory: applications in policy formation," in M. F. Kaplan and S. Schwartz (eds.). Human Judgment and Decision Processes: Applications in Problem Settings. New York: Academic Press.

HWANG, C. L. and YOON, K. (1981). Multiple Attribute Decision Making Methods and Applications: A State-of-the-Art Survey. Berlin: Springer-Verlag.

KARAKO, J. and WOLF, H. (1982). "Evaluating a street department's performance." Public Productivity Review, 6:122-127.

KEENEY, R. L. (1980). "Evaluating alternatives involving potential fatalities." Operations Research, 28:188-205.

KEENEY, R. L. (1979). "Evaluation of proposed storage sites." Operations Research, 27:48-64.

KEENEY, R. L. and NAIR, K. (1977). "Selecting nuclear power plant sites in the Pacific Northwest using decision analysis," in D. E. Bell, R. L. Keeney, and H. Raiffa (eds.). Conflicting Objectives in Decisions. New York: John Wiley, 298-322.

KEENEY, R. L. and RAIFFA, H. (1976). Decisions with Multiple Objectives: Preferences and Value Tradeoffs. New York: John Wiley.

ROHRBAUGH, J. and QUINN, R. (1980). "Evaluating the performance of public organizations: a method for developing a single index." Journal of Health and Human Resource Administration, 2:343-354.

ROHRBAUGH, J. and WEHR, P. (1978). "Judgment analysis in policy formation: a new method for improving public participation." Public Opinion Quarterly, 42:521-532.

ROOSE, J. E. and DOHERTY, M. E. (1978). "Social judgment theoretic approach to sex discrimination in faculty salaries." Organizational Behavior and Human Performance, 22:193-215.

SAATY, T. L. (1980). The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation. New York: McGraw-Hill.

SAATY, T. L. and BENNETT, J. P. (1977). "A theory of analytical hierarchies applied to political candidacy." Behavioral Science, 22:237-245.

STEWART, T. R. and GELBERD, L. (1976). "Analysis of judgment policy: a new approach for citizen participation in planning." American Institute of Planners Journal, 42:33-41.

THE OSPREY COMPANY

DEVELOPING STANDARDS FOR INTERPRETING AGENCY PERFORMANCE: AN EXPLORATION OF THREE MODELS

by

Gloria A. Grizzle

Working Paper 83-5

May 1983

accepted by Public Administration Review

This paper was supported in part by Grant Number 80-IJ-CX-0033 from the National Institute of Justice, U.S. Department of Justice. Points of view or opinions stated in this paper are those of the author and do not necessarily represent the official position or policies of the U.S. Department of Justice.

CONTINUED

1 OF 3

DEVELOPING STANDARDS FOR INTERPRETING AGENCY PERFORMANCE: AN EXPLORATION OF THREE MODELS

Organizational theory has long been concerned with the question of organizational effectiveness.[1] That organizations have multiple and sometimes conflicting goals is well known.[2] Not all these goals focus upon performance in terms of achieving the organization's mission. Building employee cohesion and morale and acquiring resources for growth are examples of goals that do not.

Quinn and Cameron have recently suggested that the criteria important in evaluating an organization's effectiveness depends upon its stage in the life cycle.[3] They suggest that there are four stages in an organization's life cycle--entrepreneurial, collectivity, formalization and control, and, finally, elaboration. For the formalization and control stage, which is the stage that most stable organizations seem to be in, they suggest that the effectiveness criteria emphasize planning, goal setting, efficiency, productivity, information management, communications, stability, and control.[4]

This paper explores statistical models that may be useful to an organization during its formalization and control stage. These models may provide useful information as a part of an organization's management control system.

Two essential elements of a management control system are (1) information about how the organization is operating and (2) standards against which to compare this information to judge how well the organization is operating.[5] While developing performance measures and collecting performance data can be a difficult, time consuming, and

expensive task, this information is not of much use to managers unless it identifies areas that need corrective action. Identifying problem areas requires comparing operations information to standards or benchmarks. Possible sources of such standards for public sector organizations include an organization's goals, objectives, or targets; standards established by relevant professional associations; the performance of similar organizations; the organization's own historical performance record; and optimal or technically efficient performance levels.[6]

This paper discusses statistical models that managers can use to develop standards based on three of these sources:

(1) the organization's own objectives;

(2) the optimal level, given specified environmental and technical constraints;

(3) the performance of other organizations.

Each section below presents a model, illustrates its usage with examples for probation agencies, and discusses the model's advantages and disadvantages.

Performance Ratio Model

The performance ratio model[7] combines many measurements to produce an overall indicator of agency performance. This model uses an agency's objectives to develop ratios of actual performance to objectives. It combines data on cost and outcomes with objectives, permitting the incorporation of both efficiency and effectiveness performance dimensions. The equation is as follows:

$$P = \sum_{i=1}^{n} W_i \frac{O_i}{G_i} \div \frac{A}{B} ,$$

where P is the indicator of overall performance, G_i is the goal or objective set for the ith performance measure, O_i is the actual performance measurement for the ith measure, W_i is the weight or importance of the ith measure relative to all other measures in the set, A is the actual total spending by the agency, and B is the agency's total budget for the period for which performance was measured.

If all measurements equal the objectives and actual spending equals the budget, then P, the overall performance indicator, will equal 1.00. If performance exceeds the objectives, P will be greater than 1.00. Similarly, if objectives exceed performance, P will be less than 1.00.

To illustrate the equation's usage, assume that a probation agency received a budget for 1983 amounting to $200,000. Of this amount, $183,000 was actually spent. Further, for the sake of brevity in developing this example, assume that the chief probation officer selects 4 measures that adequately capture the important aspects of the agency's performance. These 4 measures, the objectives for 1983, the performance measurements for 1983, and the relative importance of each measure are listed below:

Performance Measure	Performance Objective	Performance Measurement	Relative Importance
% of agency effort devoted to offender supervision	75%	60%	10%
% of offenders who successfully complete their sentences without violating their conditions of probation	50%	55%	40%
% of offenders with financial obligations who keep payments current	90%	40%	30%
% of offenders employed or otherwise socially productive fulltime	70%	80%	20%
			100%

Applying the performance ratio model to these data yields a performance indicator of .96:

$$P = \sum_{i=1}^{n} W_i \frac{O_i}{G_i} \div \frac{A}{B}$$

$$P = \left[.10\left(\frac{60}{75}\right) + .40\left(\frac{55}{50}\right) + .30\left(\frac{40}{90}\right) + .20\left(\frac{80}{70}\right)\right] \div \frac{183,000}{200,000}$$

$$P = (.08 + .44 + .13 + .23) \div .915$$

$$P = .88 \div .915$$

$$P = .96$$

Note that the first term in the equation is the agency's effectiveness when outcomes are compared with objectives. These objectives should be set for some specific funding level. If, during the course of the year, spending is held below the original budget (possibly due to freezes placed on positions or across-the-board cuts to keep spending within revised revenue projections), the second term in the equation acts to lower the level of performance expected of the agency. In doing so, the equation assumes constant returns to scale (which is probably incorrect).

One advantage of this mdoel is that its simple arithmetic permits making the calculations by hand. Another advantage is that it is easy for people who have little statistical or mathematical background to understand.

One problem in using the model is the necessity of obtaining weights for the relative importance of the various performance measures included in the equation's first term. In our illustration, the chief probation officer might sit down by himself and decide that the first measure is the least important, the second measure is 4 times as important as the first, the third measure is 3 times as important as the first, and the last

measure is twice as important as the first. Depending upon management style, the chief might instead hold a group meeting where all the agency's probation officers arrive at the weights by consensus. Many techniques have been developed for estimating such weights on a systematic basis. These techniques carry such names as the analytic hierarchy procedure, multiattribute utility theory, and social judgment theory.[8]

Another objection to the performance ratio model might be that the effectiveness and spending terms are inadequate to capture other important performance dimensions. The illustration used ignores equity in the distribution of services or penalties and the quality with which the agency carries out its activities. To respond to this objection, the effectiveness term can be broadened to include equity and quality measures. Someone must still, however, make a judgment about the relative importance of outcome, equity, and quality measures.[9]

A third problem may be the assumption of constant returns to scale. If this assumption does not provide a reasonable approximation of the relationship between different levels of resources and performance, the form of the model would need to be modified.

Linear Programming Model

As a planning tool, linear programming is not new to public administration. Examples of its use include pupil assignments to school districts, hospital manpower shift scheduling, and assigning faculty teaching schedules.[10] As a performance monitoring tool, however, linear programming can be used to develop a standard against which one can compare an agency's actual performance. Developing this standard consists of estimating the optimal level of performance, given the agency's technology and the environmental conditions within which the agency must operate.

Estimating the optimal level of performance requires knowing three things:

1. the laws, procedural regulations, and resource constraints under which the agency must operate;
2. the technology by which the agency achieves its objectives;
3. the rate at which agency activities translate into achievement of objectives.

The statistical model consists of an objective function that is to be maximized subject to a series of equations representing the agency's technology and the environmental constraints within which it must operate. An illustration follows.

Suppose that a probation agency has two major tasks--conducting pre-sentence investigations and supervising offenders placed on probation. Supervision consists of some contacts with probationers that are made in the field and other contacts that are made in the probation office. Assume the following set of constraints:

1. The agency's officer hours available for these activities in a month total 2200. The average time requirements to complete one pre-sentence investigation is 6 hours; to complete one field contact, 2 hours; and to complete 1 office contact, 1 hour.
2. The objective to maximize is the number of violation-free offender days. It has been determined that an office contact has the effect of producing 12 violation-free offender days, a pre-sentence investigation contributes 6, and a field contact contributes 30.

3. Each month an average of 100 new cases is added and 100 cases are terminated. The total average caseload is 1100. For new offenders, the first contact must be in the office, not the field. All offenders must be contacted at least once a month, either in the office or the field.

4. An average of 150 offenders are sentenced each month. Judges in the agency's jurisdiction require pre-sentence investigations on about one third of the offenders before sentencing.

Given these policy and resource constraints, what is the optimal level of performance possible for this agency? We begin by stating the objective function to be maximized:

Maximize objective attainment $= 12X_1 + 30X_2 + 6X_3$,

where X_1 = the officer hours allocated to office contacts, X_2 = the hours allocated to field contacts, and X_3 = the hours allocated to pre-sentence investigations. The coefficients are the transformation rates described in assumption 2 above. This objective function is subject to the following constraints:

a) $X_1 \geq 100$ (all the new cases must be contacted in the office the first month)

2) $X_1 + X_2 \geq 1100$ (all cases must be contacted at least once during the month)

c) $X_1 + 2X_2 + 6X_3 \leq 2200$ (the effort devoted to all three activities must not exceed 2200 hours during the month)

d) $X_3 = 50$ (one third of offenders sentenced in a month must be investigated before sentencing).

Solving this set of equations indicates that the optimal objective attainment is 27,900 and that this optimum can be achieved when the agency spends 100 hours on office contacts for new cases, 300 hours on investigations of offenders, 1600 hours on 800 field contacts, and 200 hours on 200 office contacts.

Maximum violation-free offender days $= 12X_1 + 30X_2 + 6X_3$

$= 12(300) + 30(800) + 6(50)$

$= 3600 + 24000 + 300$

$= 27,900$

By inserting the actual number of hours spent on each of the three activities into the objective function, one can estimate actual objective attainment for a given month. If the policy, resource, and technological constraints have been accurately represented in this statistical model, it would not be possible for actual performance to exceed the optimum estimated by the model. By dividing optimal attainment into actual attainment, one can calculate the actual number of violation-free offender days as a percentage of the best attainment possible, given the existing technology and policy and resource constraints.

The linear programming model makes more severe information demands than does the performance ratio model. To use this model, one must identify the important activities that contribute to attaining the agency objective, calculate the resources required to produce a single unit of each of these activities, and estimate the contribution that each unit of activity makes toward achieving the objective. Estimating the contribution that each activity makes toward achieving an objective such as violation-free offender days is not a simple matter. One empirical

method would be by using a two-stage production function.[11] In the first stage, the agency's outputs would be estimated. In the second stage, these outputs would be entered as independent variables along with other influencing variables to estimate the outcomes or objective attainment. The coefficients of these outputs from the second-stage production function could be used to estimate coefficients for the objective function used in the linear programming model.

Linear programming also makes the following assumptions:

1. Allocations of resources to activities are made under conditions of certainty.
2. Variable inputs and outputs are divisible.
3. Activities can be added together.
4. Relations between variables are proportional (i.e., constant returns to scale).

These assumptions seem to be reasonably well met in the hypothetical illustration described in this section.

As is the case for the performance ratio model, the linear programming model is not hard for the nonstatistician or nonmathematician to understand. It is, however, usually too tedious to solve by hand and is usually solved by computer.

Cost Function Model

This last model permits comparing an agency's performance with the performance of other agencies. Economists have developed cost and production functions for looking at the efficiency with which an operation transforms inputs into outputs. The two types of functions are alternative ways of looking at efficiency. Production functions define the maximum

output possible for some specified level of inputs, given existing technology. Cost functions define the minimum cost of producing outputs, given input prices and existing technology.

Both production and cost functions can be used to develop a standard against which a specific agency's performance can be compared. The function appropriate for an agency depends upon whether the agency's administrators have greater control over outputs or costs. A production function can appropriately be used when the level of output is largely under the control of agency administrators. If, on the other hand, agency administrators have greater control over costs than over outputs, the cost function is more appropriate.[12] Probation agency administrators have little control over the level of output. The courts, not the agency, have primary control over the number of offenders the agency supervises and how long the agency supervises each probationer. While probation agencies do not by any means have total control over costs, they do exercise more control over cost than over output. Therefore, this illustration uses a cost function rather than a production function.

For a cost function, the dependent variable is the total cost of operating the probation agency for a given period of time, e.g. a month or year. The form of the function that we use to estimate an agency's expected total cost is called homothetic Cobb-Douglas. This form allows costs to vary with output in rather complex ways.[13] In simplest form, cost depends upon the quantity of output and the prices of the resources required to produce that output. Output for a probation agency may be defined as the number of offenders supervised. The model assumes that all agencies being compared use the same process or technology to supervise

offenders. The major resource in a probation agency is its personnel. Other resources that might be priced and included in the equation include office space and travel costs. This simplified model is shown below:

$$\ln TC = a + b_1 P + b_2 \ln P + c_1 \ln E + c_2 \ln S + c_3 \ln T + \varepsilon$$

In this equation, ln stands for logarithm. TC is the total cost of operating the probation agency; a is the intercept; P is the output - i.e., the number of probationers supervised; E is the salary and fringe benefit price paid for personnel; S is the price paid for office space; T is the price paid for travel; and ε is the error term that represents influences on total cost that are not captured in the model. The terms b_1, b_2, c_1, c_2, and c_3 are coefficients that are estimated when the model is applied to data measuring agencies' outputs and prices.

To use this model for establishing a standard against which to compare an agency's performance, one must first collect data on outputs and prices for a number of different agencies. The model then can use these data to estimate the average total cost and the extent to which output and prices influence total cost. The coefficients in the model would then be replaced by numbers that reflect the amount of influence each output or price term has upon total cost. Once this equation has been estimated, one can use it to develop an expected total cost for an agency, given its actual output and prices. This expected total cost may be used as the standard for comparing actual total cost. If actual total cost is lower than expected total cost, one would conclude that the agency is operating more efficiently than the average for the set of agencies to which it is being compared.

As an example, assume that the actual cost for agency K was $200,000. Further, assume the model estimated agency K's expected cost (based upon the number of probationers agency K actually supervised and the prices it actually paid for personnel, office space, and travel) as $220,000. By dividing actual cost by expected cost ($200,000 ÷ 220,000 = .91), subtracting this quotient from 1.00 (1.00 - .91 = .09) and multiplying the answer by 100% (.09 X 100% = 9%), one might conclude that the agency operated 9% more efficiently than the average in its comparison group.

At this point the reader may reasonably object to using this expected cost as a standard and point to other factors that need to be taken into account. In its present form, the cost function model captures a narrower range of performance variables than did the performance ratio and linear programming models. One ought, for example, to take into account variations in quality of service and outcomes instead of simply looking at the quantity of output. Also, agencies should not be compared with each other unless differences in the characteristics of the probationers they supervise are taken into account. Finally, one might argue that the socioeconomic characteristics of the community in which the agency is located need to be taken into account.

The cost function model can be expanded to include measures of these other factors that affect agency performance. Examples of measures of service quality that might be added to the model include the percentage of probationers for whom needs assessments are completed, the percentage of referrals to helping agencies that probation officers follow up, and the number of times probation officers contact the probationers they supervise. Examples of outcome measures that might be added include the

percentage of probationers who complete their sentences without violating the conditions of their probation, the percentage of offenders with financial obligations who keep their payments current, and the percentage of probationers employed or otherwise socially productive. Probationer charactristics that might be important include the types and severity of offenses committed and the types of help needed to become productive, law-abiding citizens. Community socioeconomic characteristics that might need to be taken into account include the availability of community resources that probationers may be referred to when they need help, the availability of jobs and affordable housing, and the availability of a public transportation system for moving the probationers to and from work.

One can see that using cost functions to compare an agency's performance to the average for other agencies is more demanding in terms of the data required than the performance ratio and linear programming models. More measures are required to take into account the effects of environmental factors upon agency performance. Also, as the number of terms included in the model are increased to compare fairly an agency's performance with other agencies, data from a greater number of agencies are required to make the model workable.

While cost functions offer the advantage of comparing performance with other organizations, it is important that the organizations compared share common processes. The cost function does not identify what the technological process is but assumes that the organizations whose data are used to develop the standard share homogeneous technological processes. Hanushek believes this assumption may not appropriately characterize educational organizations,[14] and his caveats apply equally well to many other types of public organizations.

First, an organization's production technology may be changing during the measurement period. Second, even if the technology is stable, workers may have considerable discretion in the process they choose to use. Some macro organizational and process characteristics may be clearly defined and reproducible practices. Examples for educational organizations include class scheduling, curricula, and organizing teaching and research faculty into departments. Other practices, however, may vary from worker to worker. Examples for teachers include techniques used in the classroom to present material, how students are tested, and methods used to involve students in the learning process. If cost functions are used to develop performance standards, extreme care should be taken to ensure that the organizations compared to each other use the same macro and micro processes.

Another disadvantage of working with cost functions is their complexity. Once the data are collected, they must be fitted with the appropriate functional form. This part of the process requires both computers and someone with considerable statistical expertise. The amount of data required and the complexity of the statistical modeling make cost functions the most costly of the three models discussed. Yet many observers of public organizations have a keen interest in comparing performance across agencies. Given the interest in cross-agency comparisons, further research on these models seems warranted.

Summary and Conclusion

Each of the three statistical models discussed permits comparing an agency's performance to some standard or benchmark. The performance ratio model compares performance to the agency's own objectives. The linear

programming model estimates an optimal level of performance to which the agency's performance can be compared. The cost function model permits comparing the agency's performance to the average performance of other agencies with similar cost-influencing characteristics. Table 1 summarizes the salient characteristics of these models.

The performance index derived from the performance ratio model is sensitive to the levels at which the performance objectives are set. Linking future funding to how well the organization performs on this index gives organizations an incentive to set objectives at a low level so that they can be certain to attain them. One way of alleviating this problem is to have the agency negotiate objectives with those who fund the agency and hold it accountable. In the absence of mutually-agreed-upon objectives, the integrity of the performance index might be protected by using the model as an aid to internal management and not as a tool that other organizations can use to hold the agency accountable for its performance. The model's simplicity and moderate data demands make it an attractive tool provided the objectives included in the model are realistic.

When an organization can identify its own processes but is not sure that other organizations use the same processes, linear programming may be an appropriate method of developing a standard. This model requires not only process identification but also the ability to measure the rate at which each process or activity transforms resources into outputs. If this information is available, a linear programming model can estimate the optimal level of performance, which becomes the standard against which to compare actual performance. It can also be a useful planning and management tool for answering "what if" questions. By changing the

constraints, one can estimate the effect that introducing new policies or deleting policies would have upon the organization's performance. The model can generate a new standard by estimating what would be optimal performance if a policy or set of policies changed.

Finally, the cost function model is the only one of the three explored that permits comparing agency performance across agencies. This capability is obtained at the cost of more extensive data collection and more sophisticated modeling skills.

All these models can provide a standard against which to compare overall agency performance. Which model may be most appropriate for an organization to use for generating performance standards depends upon several factors:

(1) the purpose for which the organization wants to use the performance standards;

(2) whether the assumptions upon which the model is based conform to the organization's characteristics;

(3) whether the data the model requires to estimate the performance standards are available;

(4) the level of staff expertise and resources that the organization can allocate to developing performance standards.

When developing or improving their management control systems, some organizations may find it most appropriate to use their own objectives as the standard against which to compare actual performance. Others may be able to incorporate optimal levels as standards, and still others might need to compare their performance with that of similar organizations.

Table 1

Comparison of Three Models for Comparing Agency Performance

Model	Type of Comparison	Data Requirements	Difficulty of Interpreting Results	Calculation Aids Required
Performance Ratio	Agency's own objectives	Least	Easy	Paper and Pencil
Linear Programming	Optimal performance level	Intermediate	Intermediate	Computer and software
Cost Function	Similar agencies	Most	Difficult	Computer and software

Footnotes

[1]For recent literature reviews see Robert E. Quinn and John Rohrbaugh, "A Spatial Model of Effectiveness Criteria: Towards a Competing Values Approach to Organizational Analysis," Management Science, 29:3 (March 1983), pp. 363-377; Terry Connolly, Edward J. Conlon, and Stuart J. Deutsch, "Organizational Effectiveness: A Multiple Constituency Approach," Academy of Management Review, 5:2 (April 1980), pp. 211-217; Kim Cameron, "Measuring Organizational Effectiveness in Institutions of Higher Learning," Administrative Science Quarterly, 23:4 (Dec. 1978), pp. 604-632.

[2]For an excellent discussion of different goals ascribed to organizations, see Charles Perrow, "Demystifying Organizations," in Rosemary C. Sarri and Yeheskel Hasenfeld (Eds.), The Management of Human Services (New York: Columbia, 1978), pp. 105-120.

[3]Robert E. Quinn and Kim Cameron, "Organizational Life Cycles and Shifting Criteria of Effectiveness: Some Preliminary Evidence," Management Science, 29:1 (January 1983), pp. 33-51.

[4]Ibid., p. 48.

[5]Lennis M. Knighton, "An Integrated Framework for Conceptualizing Alternative Approaches to State Audit Programs" in S. Kenneth Howard and Gloria A. Grizzle (Eds.), Whatever Happened to State Budgeting? (Lexington, Ky.: Council of State Governments, 1972), pp. 162-174.

[6]Kim Cameron, "The Enigma of Organizational Effectiveness," New Directions for Program Evaluation, 11 (September 1981), 1-24; Gloria A. Grizzle, et al., Measuring Corrections Performance (Washington, D.C.: Government Printing Office, 1982); Harry P. Hatry, "Performance Measurement Principles and Techniques: an Overview for Local Government," Public Productivity Review, 4:4 (December 1980), 312-339.

[7]Adapted from Jin Eun Kim, "Cost-Effectiveness/Benefit Analysis of Postsecondary Occupational Programs: A Conceptual Framework," Planning and Changing, 11:3 (Fall, 1980), pp. 150-165.

[8]Seventeen of these methods are classified in Ching-Lai Hwang and Kwangsun Yoon, Multiple Attribute Decision Making Methods and Applications: a State-of-the-Art Survey (Berlin: Springer-Verlag, 1981).

[9]Definitions of these types of measures may be found in Gloria A. Grizzle, "A Manager's Guide to the Meaning and Uses of Performance Measurement," American Review of Public Administration, 15:1 (Spring 1981), 16-28.

[10]See, for example, Allen D. Franklin and Ernest Koenigsberg, "Computed School Assignments in a Large District," Operations Research, 21:2 (March-April, 1973), pp. 413-426; Marvin Rothstein, "Hospital Manpower Shift Scheduling by Mathematical Programming," Health Services

Research, 8 (Spring 1973), pp. 60-66; Gordon B. Harwood and Robert W. Lawless, "Optimizing Organizational Goals in Assigning Faculty Teaching Schedules," Decision Sciences, 6:3 (July, 1975), pp. 513-524.

[11]For this idea I am indebted to Ann D. Witte, Associate Professor of Economics, University of North Carolina at Chapel Hill.

[12]Ann D. Witte, "Using Production and Cost Functions to Measure the Efficiency of Corrections Agencies" (Tallahassee, Fla.: Osprey, 1982), p. 17.

[13]Gloria A. Grizzle and Ann D. Witte, "Efficiency in Corrections Agencies," in Gordon L. Whitaker and Charles Phillips (Eds.), Measuring Performance in Criminal Justice Agencies, Vol. 18, Sage Criminal Justice Systems Annuals (forthcoming, 1983).

[14]Eric A. Hanushek, "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," Journal of Human Resources, 14:3 (1979), 351-388.

THE OSPREY COMPANY

EFFICIENCY IN CORRECTIONS AGENCIES

BY

Gloria A. Grizzle

and

Ann D. Witte

accepted for Analyzing Performance in Criminal Justice Agencies

Sage Criminal Justice System Annual, Volume 18, 1983

Working Paper 83-8

August 1983

Supported in part by grant 80-IJ-CX-0033 from the National Institute of Justice, U. S. Department of Justice. Views and opinions are those of the authors and do not necessarily reflect the official position or policies of the U. S. Department of Justice. We would like to thank Peter Schmidt of Michigan State University for his aid in obtaining estimates of inefficiency for each prison in our sample.

EFFICIENCY IN CORRECTIONS AGENCIES

This chapter explores the problem of providing performance information for corrections agencies. Suggested in the sections below are an appropriate theoretical perspective from which to view an agency's performance, types of performance measures consistent with this theoretical perspective, and statistical models by which one can interpret performance. The cost and production function approach is then developed in more detail as a method of measuring an agency's efficiency. Data from prisons illustrate the use of this approach.

Theoretical Perspective

Performance measurement means obtaining information useful to someone in assessing how well an organization or program is working. What measures are relevant to this task depend upon one's theoretical perspective. Organizational theory provides a number of different models by which to assess an organization's success (Cameron, 1981; Quinn and Rohrbaugh, 1981). The principal models are listed below. In parentheses following each model type are the criteria that best conform to that model's perspective on organizational success.

Human relations model (human need satisfaction, human resource development, morale, cohesion)

Rational goal model (the organization's stated goals)

System resource model (organizational growth, resource acquisition, external support)

Internal process model (stability, smooth functioning, absence of internal strain)

Strategic constituencies model (satisfaction of important constituent groups)

Quinn and Cameron (reported in Anderson, 1981) review seven models of organizational life cycles and develop a model suggesting that organizations progress through four stages:

Entrepreneurial--early innovation and creativity

Collectivity--informal communication and structure, sense of family and cooperation among members, personalized leadership

Formalization and control--organizational stability, efficiency of production, rules and procedures

Elaboration of structure--decentralization of structure, orientation to external environment, adaptation

Except for some community-based programs, corrections organizations in the United States generally seem to be in the formalization and control stage.

Quinn and Cameron also suggest that the appropriate model by which to judge an organization's performance depends upon the organization's stage in its life cycle. They believe that the rational goal and internal process models provide the most appropriate perspectives from which to measure the success of organizations in the formalization and control stage. Of these two models, the rational goal model would seem the more useful in serving such purposes as public accountability, program planning, resource allocation, and operations analysis.

Assuming that the rational goal model is the most appropriate model for measuring corrections performance, which performance dimensions

conform to this model? Public administrationists, economists, and political scientists have all advocated performance measurement dimensions consistent with this model.

The Technique of Municipal Administration (ICMA, 1958), for example, suggested a variety of dimensions for measuring performance: costs, efforts (man-hour units), performance (work units, production), results, effectiveness, needs, and ability of the tax base to support a certain level of expenditure. Other researchers have adapted an economist's orientation and look at public sector programs as production functions that transform inputs into outputs. Bradford, Malt, and Oates (1969) expanded the classic input-program-output concept by distinguishing between outputs directly produced and the consequences of those outputs. More recently, Bahl and Burkhead (1977) added an additional component to this set of performance dimensions--environment (or needs of the citizenry).

While economists have focused primarily upon the relationship between inputs and outputs, political scientists have argued that the distribution of outputs must also be considered (Bodily, 1978; Coulter, 1980; Jones, 1981; Lineberry and Welsh, 1974; Ostrom, Parks, Percy, and Whitaker, 1979; Wilenski, 1980-81). Suggested standards for measuring equitable distribution of service include input equality, output equality, categorical equality, and demand. Recently, Bryan Jones has tied the equity concerns to the dimensions derived from the production function concept (Jones, 1981). He sees government programs as service production processes that involve four transformations:

inputs → activities → outputs → outcomes → impacts.

He argues that each step in the process has distributional effects, benefiting some people and costing others.

It is this last conceptualization of performance that seems most useful in identifying the information that should be available when assessing corrections performance. This conceptualization draws attention to these questions:

What do corrections agencies produce?

What are the benefits?

Who benefits?

Who pays?

How cost-effective are corrections agencies?

How efficiently do corrections agencies operate?

We focus the remainder of this chapter upon the question of how to measure and interpret a corrections agency's efficiency. In our discussion, we address both technical and allocative efficiency. An agency is considered technically efficient if, given some set of resources, no more output can be produced by changing the way the resources are combined (Levine, 1981). Allocative efficiency exists when one uses the production process that maximizes physical output per dollar value of input (Levine, 1981).

Bases for Comparisons

Measuring performance implies the ability to compare two pieces of data. First, one needs data that describes how an agency is operating over specific time period--a performance measurement. An example of such a measurement might be Agency A's cost per probationer supervised in 1982 ($1,000). Standing alone, this measurement does not permit one to

conclude whether $1,000 is adequate or inadequate performance. The second piece of data needed is a benchmark to which one can compare this performance measurement in order to judge how well the agency is operating. Continuing with the unit cost example, assume that lower unit costs indicate better performance and that the benchmark established is $1,300. After comparing Agency A's unit cost of $1,000 to the benchmark of $1,300, one would then conclude that its performance (as measured by unit cost) was good.

What is the source of the benchmark against which one compares a performance measurement? Possible sources include an agency's goals, objectives, or targets; standards established by relevant professional associations; the performance of other agencies; the agency's own historical performance record; and optimal or technically efficient performance levels (Cameron, 1981; Grizzle, et al., 1980; Hatry, 1980).

Decision makers above the agency level will probably want to compare performance across agencies. The great diversity of missions, programs and clientele groups among corrections agencies, however, requires that one exercise special care when comparing agencies. Performance comparisons are most appropriate when these conditions exist:

(1) When performance is measured in terms of unit cost, the agencies to be compared have common products or outputs, similar conditions under which to operate, and similar inputs.

(2) When performance is measured in terms of equity, potential clientele groups of the agencies to be compared have similar characteristics.

(3) When performance is measured in terms of outcomes, the agencies to be compared have similar outcome objectives or missions and work in similar external environments.

(4) Agencies use the same definitions, accounting methods and data collection and reduction procedures.

(5) Data collection and reduction techniques are practical and relatively cheap.

(6) Agencies have an opportunity to explain unusual situations.

(7) Timely data collection and reporting occurs.

(8) Agencies operate under similar laws and procedural regulations.

(9) Agencies operate under similar incentives for collecting and reporting performance measurements accurately.

Models for Generating a Single Measure of Overall Agency Efficiency

Much of the total effort that has gone into performance measurement research during the past twenty years has been devoted to the problem of how to combine multiple measures into a single performance measure. Three such approaches, discussed below, are multiattribute decision theory, data envelopment analysis, cost and production functions.

Multiattribute Decision Theory

Many multiattribute techniques have been developed for the purpose of combining multiple attributes or outcomes into a single measure. All these techniques develop weights for each individual attribute. These weights are used as coefficients for the attributes and, through an aggregating function, the weighted attributes are combined into a single measure. Applications of these techniques may differ in terms of

1) who identifies the important attributes
2) who sets the weights
3) how the performance of each agency for each attribute is determined
4) the aggregating (or utility) function used
5) how overall performance is calculated.

We discuss four of these techniques--decision analysis, simplified multiattribute rating technique, the analytic hierarchy procedure, and social judgment theory--and summarize an application of each.

Decision Analysis (DA).—Decision analysis may be the best known of the multiattribute methods discussed here (Keeney and Raiffa, 1976). It is also the most complicated of the methods to explain. This technique divides the performance measurement process into several tasks—identifying the dimensions against which an agency's performance shall be assessed, measuring each agency's performance in terms of each dimension, determining weights by estimating the relative importance of each dimension, and combining the performance measurements and weights so as to generate an overall performance measurement for each agency. The decision analysis application discussed here involves designing police sectors for a city (Bodily, 1978).

The purpose of this New Haven-based exercise was to determine which police sector designs would perform best, taking into account two different performance dimensions. The participants included a consultant, an administrator, a citizen representative, and a police representative. The consultant identified two performance dimensions—equality of travel time and equality of workload—and estimated each

design alternative's performance on these dimensions, using a hypercube queueing model. The administrator judged the relative importance of these two dimensions. A citizen representative specified a travel time for which he was indifferent between two sector designs. One design offered a travel time of 1 minute in the first sector and 10 minutes in the second sector. The second design offered equal travel times for the two sectors, with the equal time being that specified by the citizen representative. A police representative followed a similar procedure in determining his preferences regarding the distribution of workload across sectors.

Assuming constant inequality aversion and mutual utlity independence, the consultant then combined into a single overall utility function the citizen's preference regarding travel time equality and the policeman's preference regarding workload distribution equality with the administrator's relative weights for travel time and workload equality.

Inserting travel times and workload distribution estimates for a given sector design into this function and evaluating the function yields a single overall performance measurement for that sector design. The sector design with the highest overall measurement would be expected to perform best in terms of the two performance dimensions addressed.

Simplified Multiattribute Rating Technique (SMART).—Ward Edwards, who developed this technique, has used it to evaluate alternative desegregation plans for Los Angeles schools (Edwards, 1979). As the name implies, SMART is designed to simplify the kinds of judgments required by Keeney and Raiffa's decision analysis technique. Edwards assumes that the organization, rather than a single individual, is the decision maker.

Following this assumption, he partitions the decision problem and looks to individuals with different expertise to render judgements for different parts of the problem (Edwards, 1980).

In the Los Angeles application, school board members identified the performance dimensions in terms of which the alternative desegregation plans would be judged. They also determined the relative importance of these dimensions. Examples of dimensions selected include the plan's effect upon racial-ethnic composition of schools, educational quality, community acceptance, and stability. Edwards used a direct scaling technique to elicit each board member's judgment of the relative importance of these dimensions and then averaged the individual weights to produce a single set of weights.

School district staff then estimated how well each desegregation plan would perform in terms of each dimension. Estimates were located on a 0 to 100% scale. The utility function used was linear and additive, i.e., the utility for each plan was calculated by multiplying each outcome times its respective weight and summing the products.

Analytic Hierarchy Process.--Thomas Saaty developed this technique in 1971 and has applied it to many policy issues since that time. The application reviewed here considers the effects of seven higher education policies upon four performance dimensions (Saaty, 1980). The alternative policies evaluated were (1) the status quo, (2) vocational-technical orientation, (3) subsidized education for all, (4) education for those with money or exceptional talent, (5) all public (government owned) institutions, (6) technology based instruction, and (7) part-time teaching without research. A group of 28 college-level teachers identifed four

performance dimensions to which they believed higher education policies should contribute: prosperity, civil order, profit for industry, and perpetuation and power for industry. The relative importance of these performance dimensions was established through a two-step procedure. The teachers first reached a consensus on the relative importance between each pair of dimensions. These consensus ratings were then set in a matrix and the eigenvector of the matrix calculated. This eigenvector is an array of numbers that reflect the weight, or relative importance of each dimension (subsequently referred to as the importance vector).

Next, the teachers determined how well each policy would perform on each of the four dimensions, using a similar two-step procedure. Through pair wise comparisons, the teachers first reached a consensus on the relative degree to which each policy alternative would affect each dimension. These consensus judgments were then set in a matrix for each performance dimension and an effect vector calculated for each dimension. Each dimension's effect vector then reflects each policy's relative contribution toward that dimension.

The final step in the analytic hierarchy process was to construct a matrix consisting of the four effect vectors and to multiply this matrix by the importance vector. The resulting priority vector then contains a single number for each alternative that reflects its overall performance across all four dimensions.

Social Judgment Theory.--The three multiattribute techniques discussed above all elicit performance dimension weights directly, either from direct scaling or paired comparisons. These techniques then

multiply performance on each dimension by its respective weight and sum the resulting products to get a single overall performance measurement. Social judgment theory, in contrast, is a holistic approach that first requires people to give a single overall performance rating. The technique then infers the weights for each dimension from the overall performance ratings. The SJT application discussed below deals with the type of handgun ammunition that the Denver police should use (Hammond, 1976).

SJT broke the decision problem into several parts. The debate over which bullet to use centered upon the bullets' performance on three dimensions: (a) stopping effectiveness, (b) amount of injury, and (c) threat to bystanders. City officials and other interested people rated 30 profiles of hypothetical bullets that varied in terms of the extent to which each bullet performed across these three dimensions. These ratings were regressed against the hypothetical performance indicators to infer the relative importance or weights for each performance dimension.

In a separate process 5 ballistics experts judged the performance of 80 real but unnamed bullets on the same three performance dimensions. For each bullet, these performance scores were multiplied by the weights inferred from the earlier regression analysis and summed to provide an overall performance measurement.

Summary.—Table 1 summarizes the important characteristics of these four multiattribute techniques. Each of them offers a systematic means of eliciting judgments about the relative importance of different performance measures. The applications discussed vary in terms of who participates in identifying the dimensions against which an agency's

performance is to be measured, determining the relative importance of these dimensions, and scoring the agency's performance in terms of each of these dimensions. Who participates in these steps is not inherent in the technique used. The division of labor used in the decision analysis application discussed could, for example, be followed when applying one of the other multiattribute techniques.

If people already know the relative importance of performance dimensions, then using any one of these techniques to elicit weights may be superfluous. If people have no opinion about the relative importance of different performance measures, then none of these techniques will help. Of the four techniques, decision analysis is the most complicated and the method least likely to be tolerated by those whose judgments would be elicited. The simplified multiattribute rating technique is the simplest and the quickest, but some people doubt the appropriateness of a linear, additive aggregating function.

TABLE 1 HERE

Data Envelopment Analysis

Mathematical programming is a second group of techniques that can be used to combine multiple performance measures into a single indicator. The most recently developed technique in this area is called data envelopment analysis. As does the production function, data envelopment analysis allows comparisons of an agency's efficiency relative to that of other agencies. The technique synthesizes from the set of efficient agencies a piece-wise linear extremal production function.

Its proponents (Charnes, Cooper, and Rhodes, 1981) claim that data envelopment analysis does not have many of the limitations attributed to production functions. The method does not

(1) require specification of the functional form;

(2) require predetermined prices or weights for each input;

(3) assume differentiability of the frontier surfaces;

(4) assume that prices for the inputs and outputs are independent of their magnitude;

(5) assume an absence of capacity restraints for inputs.

Charnes and Cooper (1980) apply data envelopment analysis to data generated as a result of the Project Follow Through experiment sponsored by the federal government. For each school included in the analysis, three output measures were used:

(1) reading scores of students;

(2) mathematics scores of students;

(3) a self-esteem measure for students.

Five input measures were used:

(1) education level of the students' mothers;

(2) occupation level of students' family members;

(3) number of times parents visited the school site;

(4) time parents spent with students on school-related topics;

(5) number of teachers at each school site.

For the 49 Program Follow Through sites, relative efficiency ranged from 80% to 100%.

Production and Cost Functions

The production and cost functions developed mainly by economists are a third technique to aid in understanding the nature of operations in corrections agencies. Recent developments in the use of this technique, frontier cost and production functions, allow one to develop a single

overall measure of effectiveness. In this subsection we discuss the concepts underlying cost and production functions, emphasizing important issues to consider when studying the activities of public bodies. In the next section we illustrate the use of these techniques in understanding the operation of various types of correctional agencies.

The economic constructs of cost and production functions were originally developed to analyze the nature of production in private sector, profit maximizing firms, particularly manufacturing firms.[1] The production function summarizes, mathematically, the nature of technically efficient production. It indicates the maximum output attainable for any specified level of inputs, given the existing technology or "state of the art."

An alternative way of looking at the productive relations of the firm is through the firm's cost function. The cost function indicates the minimum costs of producing various outputs given input prices and the prevailing technology. Duality theory, which only became a well integrated part of economic production theory during the 1960s, establishes the equivalence of the cost and production function approaches to understanding the nature of firm operation.[2] The equivalence of the two approaches makes selection of one rather than the other largely a statistical and empirical issue.

In developing their constructs of cost and production functions, economists originally made a number of simplifying assumptions which have since been relaxed. First, economists originally assumed that firms produce a single, homogeneous output. Actually, most firms produce multiple outputs and the outputs produced by either a single or multi-

product firm are often of varying quality. A number of authors (Hall, 1973; Hasenkamp, 1976; Denny and Pinto, 1978) have explored both theoretically and empirically the nature of production in a multi-product firm. However, work in this area is in its infancy and, as one might expect, the production process in multi-product firms rapidly becomes quite complex.[3]

Even if a firm produces a single output, the output is rarely homogeneous. For example, although some automobile plants produce a single output, a certain model of automobile, this output is of highly variable quality (e.g., an automobile with all accessories versus one with none). Researchers have adjusted for differences in product quality by introducing a number of quality indicators into production and cost functions. In applied research many researchers have succeeded in avoiding the difficulties associated with analyzing the multi-product firm by considering relatively similar products to be a single product of varying quality. We take this approach in our work.

Economists also often find it necessary to drop the assumptions that firms seek to maximize profits. In the perfectly competitive market system beloved by economists, firms are forced to attempt to maximize profit in order to survive. A side benefit is that competition from other firms forces each firm to produce efficiently. However, when competition is not present firms are free to choose a course of action other than profit maximization.

When one moves from the profit making sector of the economy to the non-profit making and public sectors, the plausibility of profit making as a goal disappears. Consider the situation in public units such as a

police department or a prison. The public provides such bodies with funds in order to produce certain goods and services, but these public services are not sold; they are supplied to all those eligible according to some set of guidelines. For example, police services are provided to all residents of a city and correctional services are provided to all criminals with certain types of sentence. This means that, to some degree, public managers often do not control the type of product they produce or the conditions under which it is produced. Researchers studying production in public bodies suggest that the conditions of production and type of product imposed on public managers by those providing funds be controlled when estimating cost and production functions.[4]

Because public bodies that do not sell their output do not receive revenue, they cannot maximize profits even if they wish to. Rather, the public often judges such bodies by its satisfaction with the services provided for its tax dollars. While some public bodies seek to produce their services at minimum costs many others do not, but rather resort to various alternative types of bureaucratic decision making such as maximizing their budget or the size of their organizations.[5] In situations like this efficiency cannot be assumed; actual production results probably do not indicate the maximum output attainable with the given inputs, and costs observed are not the minimum attainable.

When operating in situations where efficiency is not likely to prevail, researchers using economic cost and production functions must incorporate inefficiency in their models. Quite recently economists have attempted to deal with the incorporation of inefficiency in cost and production functions. The new models developed have been called

frontier cost and production functions.[6] Frontier cost and production functions do not assume that all observed productive units are producing efficiently. Instead, they seek to infer what might be possible with efficient production by examining the performance of the most efficient firms observed. To date, two different types of frontiers have been estimated--deterministic frontiers and stochastic frontiers. Deterministic frontiers[7] assume that for the most efficient firms observed, output or costs are subject to no random effects. The observed productive relations for these firms, called the frontier firms, are used to construct the frontier cost or production function generally using linear or quadratic programming techniques. For example, a deterministic frontier cost function for the firms indicated by dots in Figure 1 would be constructed based on the experience of the firms with the lowest observed costs for each level of output. A deterministic frontier function is illustrated by the curve drawn along the lower boundary of the dots in Figure 1.

FIGURE 1 HERE

Deterministic frontiers have two major problems. First, such frontiers are extremely sensitive to outliers. For example, consider the effect of an extremely low cost firm on the frontier cost function in Figure 1. Such an outlier would pull the frontier cost function downward markedly and might well cause the frontier to reflect poorly the nature of efficient productive relationships. Second, the parameters of estimated deterministic cost and production functions have no known statistical properties. Thus, although the deterministic frontier of Figure 1 might allow us to estimate how costs vary with the level of output for the

production process under consideration, it will not allow us to say whether costs vary significantly with the level of output. Note that data envelopment analysis discussed previously was developed on the basis of the deterministic frontier work discussed above.

In contrast, stochastic frontier cost and production functions[8] allow random effects on firm output and costs to be reflected in the output frontier relationships. The essential idea behind the stochastic frontier model is that the error term is composed of two parts. A symmetric component permits random variation of the frontier across firms and captures the effects of such things as measurement error and random events outside the firms' control. A one-sided component captures the effects of inefficiency relative to the stochastic frontier. A zero value for this component indicates that the firms are effectively minimizing costs while the size of a positive value for this effect indicates the degree of inefficiency for the firms studied. Figure 2 illustrates the nature of the stochastic average cost frontier for three hypothetical prisons. Note that the stochastic frontier varies from prison to prison (for example, due to riots, fires), but has the same shape for all prisons.

FIGURE 2 HERE

A stochastic frontier allows one to estimate the mean inefficiency and the mean cost of this inefficiency for all firms studied. Such estimates of inefficiency and its associated costs can be very useful both for correctional managers and executive and legislative groups which oversee their activities. For example, estimates of the costs of inefficiency may provide managers with incentives to pursue improvements. In addition, an estimate of these costs may allow executive and

legislative organizations to set more realistic budgets. Further, by studying the factors related to relatively efficient production, managers and others may be able to suggest improvements likely to increase systemwide efficiency.

Recent work on stochastic frontier functions (Schmidt and Lovell, 1979; and Jondrow, et al., 1982) allows one to obtain efficiency estimates for individual productive units as well as the overall measure of inefficiency discussed above. Information of this type is, of course, extremely useful for management decisions, as it allows the manager to compare the relative performance of the units he or she supervises.

Application of Cost Function Approaches to Corrections

In this section we illustrate the use of cost functions to analyze the operation of correctional agencies. We begin by developing a model of the nature of production in these agencies.

A Model of Production in Correctional Agencies

The development of a model of production for correctional agencies requires the researcher to make a number of simplifying assumptions and decisions. We begin by discussing the role of correctional agencies in the criminal justice system and the way in which agency operation is affected by other executive branch agencies, the legislative branch and judiciary.

As an integral part of the criminal justice system, correctional agencies in cooperation with the police and courts are charged with preventing crime when possible and punishing the offender when it does occur. Both of these services must be performed within legally defined constraints regarding due process and humane treatment. While society sees the criminal justice system as an entity with definite goals, most

people familiar with the "system" know it to be composed of distinct agencies with only limited interaction and coordination. Indeed, ultimate decision making authority for many of the agencies rests with different levels and branches of government. Although the actions of the agencies of each segment of the criminal justice system affect the agencies of the other segments, these effects are rarely considered when the administrators of a particular agency make decisions. There is no single administrator responsible for the efficient operation of the entire system and, thus it is not possible to think of the system as a productive unit seeking to produce "justice" effectively. From an economist's perspective, we would not seek to estimate production and cost functions for the system as a whole, at least as the system is currently structured in the United States. However, in seeking to model the nature of these relationships for the individual segments of the system, it is important to understand the effect of segments other than that analyzed.

A number of researchers, mainly from operations research backgrounds,[9] have described the way in which police, courts and correctional agencies interact. Figure 3 depicts these interactions emphasizing the role of correctional agencies. The relationships among the components of the criminal justice system consist of a flow of individuals through the system, court surveillance of police and correctional activities, and the provision of evidence and other information by police and correctional agencies. As can be seen in Figure 3, interactions between agencies are many and complex, and the number of alternative paths open to offenders is large. Managers in most agencies of the criminal justice system are relatively free (baring court intervention) to make day to day (short run)

decisions about the way in which their organization will operate. For example, the administrators of a state prison system are relatively free to decide upon the housing and care of inmates. Long run decisions concerning the purchase of new capital equipment and buildings, and major changes in methods of operation (technology) are subject to both legislative and judicial review, but are largely made by agency managers or other executive branch personnel. Again to use an example from corrections, correctional managers usually propose capital improvement plans to executive budget offices. These offices in turn decide which improvements to recommend to the legislative body. The legislative body finally decides whether funds will be provided for the proposed improvements.

FIGURE 3 HERE

Having set correctional agencies in perspective and decided that correctional administrators have substantial decision making power, we need next to determine the goals which correctional agencies pursue. Recall that we perceive correctional agencies as in the formalization and control stage of organizational development. Thus, formal goals and internal processes are of primary importance when considering performance.

We, as a society, fund correctional agencies to impose restrictions on individual freedom as a punishment, to protect society by incapacitating the offender and to deter both offenders (specific deterrence) and others (general deterrence) from committing future offenses. We also often provide services to the offender, and require that they productively occupy their time. We do this for a number of reasons. First, many

believe that some services and some types of work serve to rehabilitate offenders. Second, work and other activity requirements often serve to lower the costs of correctional agencies. They may do this indirectly as well as directly. Directly they can lower the costs of operating correctional agencies by direct payment (for example, Florida requires payments from probationers and parolees), providing maintenance and other needed activities. Indirectly by occupying offenders' time, these activities may serve to lower the level of correctional supervision which it is necessary to maintain. Third, at least some members of our society believe that the provision of such activities is morally right because deprivation of liberty provides sufficient punishment and retribution for crime. Increasingly in recent years, correctional agencies have been asked to provide for and manage the payment by offender of restitution to victims.

Finally, we would like all of these goals achieved as cheaply as possible and under conditions we deem acceptable. Given this menu of often conflicting goals it is not surprising that our correctional system has often seemed to lack direction or to change directions at fairly frequent intervals. Which correctional goals are emphasized has changed through time and at any given time varies from correctional agency to correctional agency.[10] Currently, the major goal of the majority of correctional agencies appears to be control although rehabilitative programming continues.[11]

As noted above, most economic models of cost and production assume that costs are minimized. The above discussion should make it clear that cost minimization is, at best, one of many goals pursued by

correctional administrators. Pressures on correctional administrators to minimize costs generally come from outside the correctional system. For the federal correctional system, the pressures come primarily, in the executive branch, from the Office of Management and Budget; and, in the legislative branch, from congressional committees and the General Accounting Office. For state and local correctional agencies, pressures to minimize cost come from the department of finance or budget, speaking for the chief executive, and various analysis offices and staff, speaking for the legislative body.

Our work in prisons and parole and probation offices leads us to conclude that assuming that correctional administrators minimize costs is tenuous, at best. Yet, use of duality theory and standard statistical techniques (e.g., ordinary or generalized least squares) requires this assumption. We attempt to deal with this dilemma in two ways. First, when interpreting results obtained using standard estimation techniques, we are careful to point out results which depend critically on the assumption that costs are minimized. Second, we estimate frontier cost functions which explicitly model the fact that administrators may not minimize costs. Briefly to preview our results, we find that administrators do not minimize costs and, thus, our results which use traditional estimation techniques are best interpreted as behavioral equations rather than the cost curves of microeconomic theory. In the absence of cost minimizing behavior, inferences about the nature of the production function from these results requires the strong assumption that the inefficient behavior present merely neutrally transforms (i.e., shifts the cost curve up equally everywhere) the cost curve. However, we feel

that the "behavioral" cost curves estimated are in many cases as important as the cost curves of economic theory. They provide insight concerning the actual behavior of correctional agencies and, thus, may be at least as useful as economic cost curves in understanding the nature of production in these agencies and in projecting likely future behavior.

Our next task is to determine the unit we wish to analyze (the productive unit) and the output which this unit produces. We have argued elsewhere (Witte, et al., 1979) that the individual prison is the entity which most closely approximates the economic concept of a "productive unit" within the prison system.

Having decided on the productive unit, we must next decide what this unit is producing. As noted above, the dominant correctional goal currently appears to be controlling convicted offenders. Conforming to this outlook we see the output of the correctional agency as a certain number of convicted offenders supervised for a certain period of time. However, we do not see this output as homogeneous, but rather seek to introduce a number of factors which will reflect different "qualities" and types of supervision. We also recognize that different correctional agencies work in different environments and with differing types of offender.

Having developed a basic model, we must next decide whether to estimate this model using a cost or production function approach. As noted earlier this is largely a statistical and practical issue. One should estimate a production function if one believes that the level of output is largely under the control of managers (i.e., endogenous to the model) and a cost function if one believes that costs are under greater control of

managers than output. Recall that our measure of output for the correctional agency is the number of convicted offenders supervised during any given period of time, which we will refer to as the case load for the sake of brevity. Correctional managers have little control over either the size or the composition of their case load because they are required to accept all convicted offenders directed to them by the courts or other agencies. Further, correctional managers have only limited control over the release of offenders from their supervision. While costs are not entirely within the control of correctional administrators, they are to a far greater degree than output, particularly in the long run. Thus, we choose to estimate cost rather than production functions for the correctional agencies we study. We began by estimating cost functions of the following general form for selected prisons, and probation and parole offices:

$$\ln AC = a + b_1 y + b_2 \ln y + (\ln P)\acute{\Upsilon} + \acute{A}N + \acute{S}U + \acute{Q}F + \acute{\varepsilon} \quad (1)$$

where AC is the average cost of operating the correctional agency; y is a measure of the number of convicted offenders supervised; $\ln P'$ is the transpose of a vector of the natural logarithm of factor prices; a, b_1 and b_2 are parameters to be estimated; Υ, N, U, and F are vectors of parameters to be estimated; A', Q', and S' are the transpose of vectors of measures of output quality, input quality and the service conditions under which the agency operates; and ε is a vector of "disturbance terms" (or "error term") which represents random influences on average cost which we are unable to capture in our model.

One aspect of this model deserves comment. The mathematical term in which the output variable and factor prices are entered was

dictated by our choice of a homothetic Cobb-Douglas production function to represent the operations of correctional agencies. We selected this form over other alternatives because we felt that while relatively simple, it imposed important technical restrictions (e.g., diminishing marginal physical product for inputs). It also allows costs to vary with output in rather complex ways.

The Data

The data set contains information on Federal Correctional Institutions and was obtained from a number of different sources within the Federal Bureau of Prisons (FBOP), and the U.S. Department of Justice's System Design and Development Group. Federal Correction Institutions (FCIs) are generally the more modern and relatively smaller (as compared to Federal penitentiaries) medium custody institutions in the Federal prison system. FCIs hold the bulk of Federal prisoners and administrators of the Federal prison system are committed to replacing most penitentiaries with FCIs. Further, given FBOP's role as "a model for state prison systems," FCIs are likely to be a type of facility which is much utilized in the future. Appendix 3 of Schmidt and Witte (1983) contains a detailed description of this data set. It includes monthly data for the period October, 1975 through June, 1978, for all 21 FCIs that were operating.

Given our model specification (see equation 1) and data set, we can now specify empirical measures for our theoretical constructs. Table 2 summarizes our choices.

TABLE 2 HERE

Empirical Results for Prisons

We began our work by estimating a short run cost function for each of the 21 federal correctional institutions, using monthly time series data.

The dependent and independent variables used are defined in Table 2. The results we obtain, which are reported elsewhere (see Witte, et al., 1979 and Schmidt and Witte, 1983), indicate that methods of operation at the 21 FCIs varied substantially. When conducting an economic analysis of costs this means that we would not be justified in estimating a long run cost function by pooling data for all institutions. Economic theory indicates that we can only learn important facts about a particular method of operation if we study groups of facilities which are using the same method.

We searched among the 21 FCIs for a group of prisons which appeared to be using broadly similar methods of operation and were able to identify six institutions.[12] We began by using ordinary least squares to estimate our long run prison cost function using quarterly data for these six institutions for the period beginning in the first quarter of 1976 and ending in the second quarter.[13] Results are reported in the second and third column of Table 3.

TABLE 3 HERE

We are able to explain a large portion of the variation in average costs with our model, 87%. However, the coefficients of few variables are statistically significant due to extensive multicollinearity, and relatively low variance in a number of independent variables.

Desiring to rid our average cost curve specification of variables unrelated to costs, we selected two basic rules for reducing our specification. First, we retain the output and factor prices variables regardless of the significance of the coefficients on these variables as both economic theory and intuition provides strong support for their

inclusion. Second, we proceeded to sequentially drop other independent variables, beginning with the variables whose coefficient had the smallest t-ratio, until the coefficients associated with all remaining variables were significant at the 0.05 level of significance. We tested to see if we could accept the hypothesis that all deleted variables when combined were insignificantly related to average costs, and were able to accept this hypothesis. Results for the reduced specification appear in the fourth and fifth columns of Table 3.

Our work to this point has proceeded on the assumption that the correctional agencies which we study are effectively minimizing costs. Both our own work with correctional agencies and the work of others with other public and private entities lead us to believe that this is not likely to be the case. To test this assumption and to develop an overall measure of performance, we now estimate a stochastic frontier cost function using a method developed by Schmidt and Lovell (1979).

Our specification for the frontier long run average cost function is identical to our reduced specification in Table 3 above except that we reduce the specification by one variable by imposing the restriction that the factor shares sum to one. We do this to conserve both degrees of freedom and computational costs.

Specifically we estimate the following function:

$$\ln(AC) - LNCOST\text{-}L = B_0 + B_1\ LNCD\text{-}ALL + B_2\ CD\text{-}ALL + B_3\ (LNCOST\text{-}C\text{-}LNCOST\text{-}L) + A'N + S'U + \Psi \quad (2)$$

where A' is the transpose of the vector of output and input quality measures in the final specification of Table 3 and S' is the transpose of

the service condition variable in that specification. The random disturbance, ψ, is composed of two parts. One part is normally distributed with zero mean and variance σ_v^2. The other is a non-negative, half normal, random variable with a positive mean and variance σ_μ^2.

TABLE 3 HERE

The normal portion of the disturbance term captures random variations in costs between prisons that are due to factors, such as weather, riot, and fires, that are outside the individual prison decision maker's control. The half normal portion of the disturbance reflects inefficiency. This portion of the disturbance is either zero or a positive number. A zero value for this variable indicates that the prison is operating efficiently, i.e., it is a frontier prison. The size of a positive value for this variable indicates the degree of inefficiency.

We estimate our frontier cost function using maximum likelihood techniques. Results are reported in Table 4. Our frontier estimates indicate that costs will be at a minimum when the prison contains 1467 inmates.

TABLE 4 HERE

The most interesting results of the frontier estimates are our estimates for the variances of two parts of the disturbance term. Note that the variance of the half normal portion of this disturbance is quite large, indicating large differences in efficiency among prisons. We estimate that, on the average, costs in the six FCIs studied were 9.4 percent more than they would have been if the most efficient methods of operation had been utilized. Given Federal Bureau of Prison expected outlays of $327 million in fiscal year 1980, our results indicate that

efficient operation could have saved approximately $30 million. While savings of this size seem unlikely to be realized, we do believe that our results indicate that some savings may be possible.

Using a technique recently developed by Jondrow, et al. (1982), we can also estimate one extent of inefficiency at each prison in our sample for each quarter. Table 5 contains those results. Note that there is considerable variation through time in the level of inefficiency at any given prison. Indeed the temporal variation in efficiency appears to be greater than the variation across prison units. The mean estimated inefficiency for the different prison units only ranges from 8.5 percent for Prison One to 10.2 percent for Prison Six.

TABLE 5 HERE

Briefly summarizing, our analyses indicate that the average costs of incarcerating offenders at first decreases and then increases as prison population rises. According to our model, costs will be lowest when prisons are quite large (say 1000 to 1500 inmates), but not behemoth. We find further that the cost of operating these FCIs will be higher the higher is the relative cost of capital, the lower the proportion of female inmates, the older the average age of staff, and if inmates are housed in relatively large single cells, but have limited sanitary facilities.

As production and cost analysis appeared to produce quite reasonable results for the six FCIs studied, we proceeded to estimate a frontier cost function for these facilities. This function indicated that the six FCIs in the sample were not operating as efficiently as they might. As a result, costs at these FCIs were nine percent higher than they might have been. We next estimated the level of inefficiency at each prison for

each quarter. We found that inefficiency was more variable temporally than across prisons. Our results suggest that Prison One was most efficient (estimated average inefficiency 10.2 percent).

Summary and Conclusion

All three methods--multiattribute decision theory, data envelopment analysis, and production and cost functions--can be used to generate a single overall measure of agency efficiency. Data envelopment analysis and production and cost functions should probably be applied only to agencies having well defined processes, for the reasons given below. Further these two types of analyses generally provide estimates of only the efficiency aspects of agency performance while multi attribute decision theory allows one to consider other aspects of agency performance such as equity as well.

In the context of our work, production and cost theory provide much more useful guidance when analyzing the performance of large scale prisons than when analyzing the performance of probation and parole agencies. In a related effort (Witte, 1982), we estimated cost functions for five probation and parole offices. The data set contained monthly information for a single calendar year. Production and cost analysis poorly described the operation of these offices.

As a whole the work reported in this analysis of production and cost functions tends to support the conclusions of a number of other researchers who have analyzed the performance of other types of public agencies.[14] These researchers find that the economic constructs of production and cost functions are most directly applicable to public agencies which produce physical outputs (e.g., water, electricity, refuse

collection) with well defined inputs and known technological processes. For other types of public agencies, production and cost functions mainly provide useful insights as to important variables to consider and possible functional forms to be used in analysis. Production and cost analysis appears to provide fewest insights for public agencies which produce services requiring extensive interaction between public employees and the individuals receiving services (i.e., education and other social services). In such situations individual skills are extremely important and the exact way in which the service is provided may vary substantially from employee to employee.

When processes either are not well understood or vary as a result of employee discretion, multiattribute decision theory may be a more appropriate aid when generating an overall measure of agency performance. Further, multiattribute techniques can allow the researcher to relatively easily consider aspects of agency performance other than efficiency. However, multiattribute decision analysis, unlike cost and production function approaches, provides little information concerning the way in which agencies operate.

FOOTNOTES

[1]See Fuss, McFadden and Mundlak (1978, pp. 267-268) for a brief survey of recent work.

[2]For a survey of duality theory, see McFadden (1978).

[3]Some second order approximations generalize readily to multiple output. Darrough and Heineke (1978) have estimated a multiple-output translog cost function for police services. Many exact function forms, however, are intrinsically non-linear, making estimation difficult and expensive. In addition, for both exact functional forms and second order approximations, the number of parameters to be estimated for multi-product production processes quickly becomes very large if extremely restrictive assumptions are not made.

[4]See Alesch and Dougharty (1971), Hirsch (1973) or Vernez (1976) for surveys of early work and Witte (1980) for a discussion of more recent work in six areas (education, fire protection, hospitals, libraries, police protection and large scale prisons).

[5]Orzechowski (1977) provides a review of this literature.

[6]Such functions were first developed by Farrell (1957). More recent work has been done by Aigner and Chu (1968), Timmer (1971), and Aigner, Lovell and Schmidt (1977).

[7]See Farrell (1957), Aigner and Chu (1968) and Seitz (1970) for examples. Carlson (1972) estimates deterministic frontiers for higher educational institutions.

[8]See Aigner, Amemiya and Poirer (1976); Aigner, Lovell and Schmidt (1977); or Schmidt and Lovell (1979). The May 1980 issue of the JOURNAL OF ECONOMETRICS is devoted to the specification and

estimation of frontier production, profit and cost functions. The lead article by Forsund, Lovell and Schmidt (1980) contains a survey of frontier work.

[9]For an example of this work see Blumstein (1975), Chaiken, et al. (1976), and Blumstein and Larson (1976) provide an extensive survey of criminal justice models.

[10]See Martin, Sechrest and Rodner (1981) for a survey.

[11]See Minerva (1982) for a discussion of selected parole and probation offices. Witte et al. (1979) discuss the goals of large scale prisons and provide detailed analyses of the Federal and California prison systems.

[12]These institutions are Ashland, Lompoc, Lexington, Oxford, Texarkana, and Alderson. We did two things to determine if these six institutions were using similar methods of operation. First, we conducted a generalized Chow test to determine if we could accept the hypothesis that the coefficients on all variables in the short run cost function were equal across the six institutions. The value of the test statistic which is distributed $F_{125,48}$ under the null hypothesis, was 1.979. We next ran three sets of simply specified long run cost functions for subsets of these six FCIs. Specifically, we estimated cost curves for (1) the older vs. newer prisons in the group, (2) the more vs. the less secure prisons in the group, and (3) the bigger vs. smaller institutions in the group. In each case we accepted the null hypothesis that the six prisons were using similar methods of operation.

[13]Appropriate statistical tests were performed to ensure that this pooling of time series data was justified. We pooled data around the third

quarter of 1977 as detailed descriptions of the capital stock were available for that quarter when a complete physical plant inventory was conducted. The test statistic for the appropriateness of this time series pooling, which is distributed $F_{45,10}$ under the null hypothesis that pooling is appropriate, was 0.794.

[14]For example, see Alesch and Dougharty (1971), Hanushek (1979), Summers and Wolfe (1977), Vernez (1976) and Witte (1980).

Aigner, D., C. A. K. Lovell, and P. Schmidt. (1977) "Formulation and estimation of stochastic frontier production function models." Journal of Econometrics 5:1-17.

Aigner, D. J., T. Amemiya, and D. J. Poirer (1976) "On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function." International Economic Review 17:377-396.

Aigner, D. J. and S. F. Chu (1968) "On estimating the industry production function." American Economic Review 58, 4:826-839.

Alesch, D. J. and L. A. Dougharty (1971) Economies of Scale Analysis in State and Local Government. Santa Monica, California: Rand, R-748-CIR.

Anderson, D. F. (1981) "A system dynamic view of the competing values approach to organizational life cycles." Public Productivity Review 5, 2:160-187.

Bahl, R. W. and Burkhead, J. (1977) "Productivity and the measurement of public output" in C. H. Levine, ed. Managing Human Resources: A challenge to Urban Governments. Urban Affairs Annual Reviews, Vol. 13.

Blumstein, A. (1975) "A model to aid in planning for the total criminal justice system," in L. Oberlander (ed.) Quantitative Tools for Criminal Justice Planning. Washington, D.C.: Law Enforcement Assistance Administration.

Blumstein, A. and R. Larson (1976) "Models of a total criminal justice system," in L. R. McPheters and W. B. Strange (eds.) The Economics of Crime and Law Enforcement. Springfield, Ill.: Charles C. Thomas.

Bodily, S. E. (1978) "Police sector design incorporating preferences of interest groups for equality and efficiency." Management Science 24:1302-13.

Bradford, D. F.; Malt, R. A.; and Oates, W. E. (1969) "The rising cost of local public services: some evidence and reflections." National Tax Journal 22, 2:185-202.

Cameron, K. (1981) "The enigma of organizational effectiveness." New Directions for Program Evaluation 11:1-13.

Carlson, Daryl (1972) "The production and cost behavior of higher educational institutions," Paper P-36 of the Ford Foundation Program for Research in University Administration, University of California, Berkeley, California.

Charnes, A. and W. W. Cooper (1980) "Auditing and accounting for program efficiency and management efficiency in not-for-profit entities." Accounting, Organizations, and Society 5, 1:87-107.

Charnes, A.; W. W. Cooper; and E. Rhodes (1981) "Evaluating program and managerial efficiency: an application of data envelopment analysis to program fellow through." Management Science 27,6:668-697.

Chaiken, J., et al. (1976) Criminal Justice Models: An Overview. Washington, D.C.: Government Printing Office.

Coulter, P. B. (1980) "Measuring the inequity of urban public services: a methodological discussion with applications." Policy Studies Journal 8, 5:683-697.

Darrough and J. M. Heineke (1978) "The multi-output translog production cost function: The case of law enforcement agencies," pp. 259-302, in J. M. Heineke (ed.) Economic Models of Criminal Behavior. Amsterdam: North-Holland.

Denny, Michael and Cheryl Pinto (1978) "An aggregate model with multi-product technologies," pp. 249-267, in Melvyn Fuss and Daniel McFadden (eds.) Production Economics: A Dual Approach to Theory and Applications, vol. I. Amsterdam: North-Holland.

Edwards, W. (1980) "Multiattribute utility for evaluation: structures, uses, and problems" in M. W. Klein and K. S. Teilmann, eds. Handbook of Criminal Justice Evaluation. Beverly Hills: Sage, 177-215.

Edwards, W. (1979) "Multiattribute utility measurement: evaluating desegregation plans in a highly political context" in R. E. Perloff, ed. Evaluator Interventions: Pros and Cons. Beverly Hills: Sage, 13-54.

Farrell, M. J. (1957) "The measurement of production efficiency." Journal of the Royal Statistical Society, Series A, General, 120, Part 3:253-281.

Fuss, Melvyn, Daniel McFadden and Yair Mundlak (1977) "A survey of functional forms in the economic analysis of production," pp. 219-268, in Melvyn Fuss and Daniel McFadden (eds.) Production Economics: A Dual Approach to Theory and Applications, vol. 1. Amsterdam: North-Holland.

Grizzle, G. A. et al. (1982) Basic Issues in Corrections Performance. Washington, D.C.: National Institute of Justice.

Hall, Robert E. (1973) "The specification of technology with several kinds of output," Journal of Political Economy 31:878-892.

Hammond, K. R. (1976) "Externalizing the parameters of quasirational thought" in M. Zeleny, ed. Multiple Criteria Decision Making, Kyoto 1975. Berlin: Springer-Verlag.

Hanushek, Eric A. (1979) "Conceptual and empirical issues in the estimation of educational production functions," Journal of Human Resources 14, 3:351-388.

Hasenkamp, Georg (1976) Specification and estimation of multiple-output production functions. Berlin: Springer-Verlag.

Hatry, H. P. (1980) "Performance measurement principles and techniques: an overview for local government." Public Productivity Review 4, 4:312-339.

Hirsch, Werner Z. (1973) Urban Economic Analysis. New York: McGraw Hill.

Jondrow, J., et al. (1982) "On the estimation of technical inefficiency in the stochastic frontier production function model." Journal of Econometrics, 19:233-238.

Jones, B. D. (1981) "Assessing the products of government: what gets distributed? Policy Studies Journal 9, 7:963-971.

Keeney, R. L. and Raiffa, H. (1976) Decisions with Multiple Objectives: Preferences and Value Tradeoffs. New York: John Wiley.

Levine, V. (1981) "The role of outcomes in cost-benefit evaluation." New Directions for Program Evaluation 9:21-40.

Lineberry, R. L. and Welch, R. E. (1974) "Who gets what: measuring the distribution of urban public services." Social Science Quarterly 54, 4:700-712.

Martin, Susan B., Lee B. Sechrest and Robin Redner (eds.) (1981) New Directions in the Rehabilitation of Criminal Offenders. Washington, D.C.: National Academy Press.

McFadden, Daniel (1978) "Cost, revenue and profit functions," pp. 3-109, in Melvyn Fuss and Daniel McFadden (eds.) Production Economics: A Dual Approach to Theory and Applications, vol. 1. Amsterdam: North-Holland.

Minerva, Karen S. (1982) "Probation/Parole Operations," Working Paper 82-1, The Osprey Company, Tallahassee, Fl.

Orzechowski, W. (1977) "Economic models of bureaucracy: survey, extensions, and evidence" in T. E. Borcherding (ed.), Budgets and Bureaucrats: the Sources of Government Growth. Durham, N.C.: Duke University Press.

Ostrom, E., Parks, R. B.; Percy, S. L.; and Whitaker, G. P. (1979) "Evaluating police organization." Public Productivity Review 3, 3:3-27.

Quinn, R. E. and Rohrbaugh, J. (1981) "Competing values approach to organizational effectiveness." Public Productivity Review, 5, 2:122-140.

Saaty, T. L. (1980) The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation. New York: McGraw-Hill.

Schmidt, Peter and C. A. K. Lovell (1979) "Estimating technical and allocative inefficiency relative to stochastic production and cost frontiers." Journal of Econometrics 9:343-366.

Schmidt, Peter and Ann D. Witte (1983) The Economics of Crime: Applications, Theory and Methods. New York: Academic Press, forthcoming.

Seitz, W. D. (1970) "The measurement of efficiency relative to a frontier production function." American Journal of Agricultural Economics 52, 4:505-511.

Saaty, T. L. (1980) The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation. New York: McGraw-Hill.

Summers, Anita S. and Barbara L. Wolfe (1977) "Do schools make a difference?" American Economic Review 67:639-652.

The Technique of Municipal Administration (1958) 4th ed. Chicago: International City Managers' Association, pp. 352-354.

Timmer, C. P. (1971) "Using a probabilistic frontier production function to measure technical efficiency." Journal of Political Economy 79, 4:776-794.

Vernez, George (1976) Delivery of urban public services: production, cost and demand functions, and determinants of public expenditure for fire, police and sanitation services. Santa Monica, California: Rand, P-5659.

Wilenski, P. (1980-81) "Efficiency or equity: competing values in administrative reform." Policy Studies Journal 9, 8:1239-1249.

Witte, Ann D. (1980) "Economies of public service delivery systems." Summary of round table discussion of the Special National Workshop on Research Methodology and Criminal Justice Program Evaluation, Baltimore, Maryland, March 18.

Witte, Ann D., et al. (1979) "Empirical investigations of correctional cost functions." Final Report to the National Institute of Law Enforcement and Criminal Justice on LEAA Grant No. 78-NI-AX-0059.

Witte, Ann D. (1982) "Using Production and Cost Functions to Measure the Efficiency of Corrections Agencies," Working Paper 82-9, The Osprey Company, Tallahassee, Fla.

Figure 1

Illustration of a Frontier Cost Function

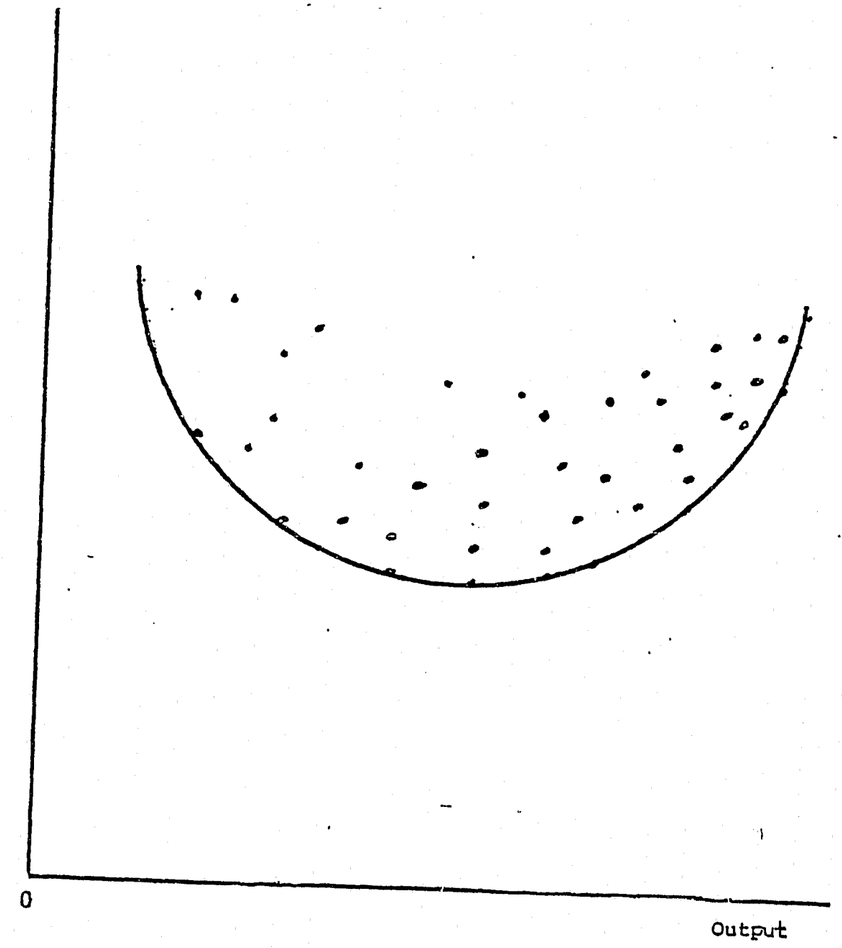


Figure 2

An Illustration of the Stochastic Frontier

Average Cost Function

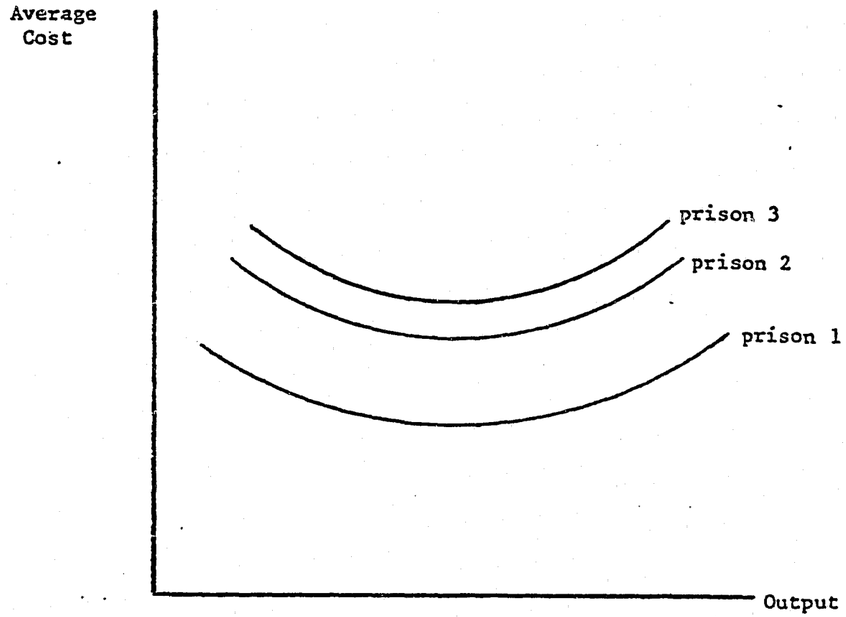
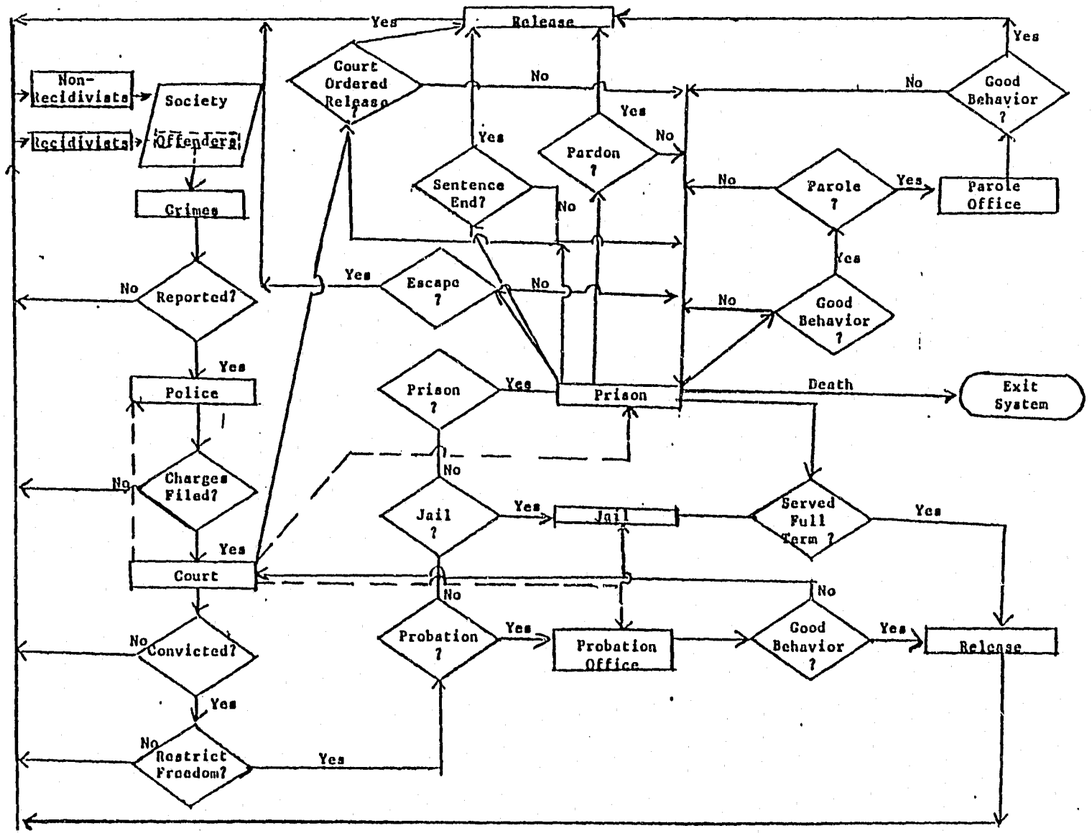


Figure 3

Flow of Offenders and Organization Interactions



Source: The left side of this diagram was adapted from Blumstein and Larson (1976, p. 473).

TABLE 1

COMPARING SEVERAL MULTIATTRIBUTE TECHNIQUES FOR MEASURING PERFORMANCE

Method	Policy Application	Who Identified Performance Dimensions	How Dimension Weights Were Set	How Each Alternative's Effect on Each Dimension Was Determined	Aggregating Function	Calculation of Overall Performance
DA	Police sector design	Consultant	Direct tradeoff by administrator	Consultant estimates based upon hypercube queueing model	Curve drawing derived from preferences of a citizen representative and a police representative using the lottery technique	For each plan, insert effect for each dimension into function and evaluate the function to calculate overall utility
SMART	School desegregation	School board members	Average of individual school board members' weights based on direct scaling of each attribute's importance	School district staff estimates, located on a 0 to 100% effectiveness scale	Linear, additive	For each plan, multiply effectiveness score for each attribute by that attribute's weight and sum the resulting products
AHP	Higher education	College-level teachers	Priority eigenvector developed from consensus reached on paired comparisons by teachers	Teachers, by consensus, assigned scores based upon paired comparisons	Linear, additive	Multiply matrix of alternatives' effect on dimensions by weight vector
SJT	Police handgun ammunition	City officials and other interested groups	Weights derived from regression analysis of ratings by city councilmen and other interested groups	Judgments from ballistics experts	Curve drawing derived from multiple regression analysis of ratings	For each bullet, insert its score on each dimension into regression equation to calculate its overall performance rating.

TABLE 2

VARIABLE DEFINITIONS AND EXPECTED SIGN OF COEFFICIENTS

Theoretical Variable and Symbol	Empirical Measure and Acronym for Prison data set	Expected Sign of Coefficient
Dependent Variable		
Costs (TC)	The sum of actual disbursements, increments, in accounts payable and non-funded costs, changes in applied costs and normal depreciation during the period divided by the number of confined days(AC)	N.A.
Independent Variable		
Output (Y)	The number of offenders incarcerated times the number of days confined during the quarter and its logarithm (CD-ALL, LNCD-ALL)	? ?
Input Prices (P)	The logarithm of the cost of capital proxied by a regional index of construction wages (LNCOST-C); the logarithm of the cost of labor, proxied by average hourly wage and fringe benefits paid to institutional staff (LNCOST-L)	+, +
Product Quality (members of the vector A)		
Security	The ratio of correctional officers to average confined population (SECURE)	+, +
Incidents	The sum of institutional escapes, inmate assaults, and violent inmate deaths (INCDNT)	+
Crowding; Deviations Short Run from Planned Output	The ratio of average confined population to institutional physical capacity and its squared value (CROWD, CROWD2)	-, +
Service Condition (member of the vector S)		
Racial Balance	The ratio of the percent non-white in the correctional staff to the percent non-white in the inmate population and its squared value (R-BAL, R-BAL2)	-, +
Auxiliary Facilities	Percent of confined days output produced in an associated camp, female facility, or detention center (PC-OTH)	?
Labor Quality		
Staff Type	The ratio of guards to other staff. (RATIO-S)	?
Education	Average years of education (ED-S)	-
Race	The percent of staff that are non-white (RACE-S)	?
Age	The average age of the staff (AGE-S)	?
Sex	The percent of staff that are male (SEX-S)	?

Table 2 (continued)

Capital Quality		
Living Area	Square feet of living area per bed (SQFPER)	?
Single Beds	Proportion of design capacity housed in single bed cells or rooms (SINGLE)	?
Sanitary Facilities	Number of toilets and urinals per design capacity (SANPER)	?
Production Quality and Service Conditions (members of the vectors A and S) Age	The average age of the inmate population in months and its squared value (AGE-I, AGE-I2)	+, -
Racial Composition	The percent of the inmate population whose race is non-white (RACE-I)	?
Sexual Composition	The percent of the inmate population whose sex is female (FEMALE)	?
Occupation	The percent of the inmate population whose longest job prior to incarceration was professional, technical, managerial, or in accounting (WCOLLAR)	?
IQ	The average Beta IQ of the inmate population (BETAIQ)	?
Sentence	Average length, in years, of the sentences of the confined population (LENGTH)	?
Crime Type of Offender	Percent of the confined population sentenced for a crime against a person (O-PERS); percent of the confined population sentenced for property offenses (O-PROP)	+, ?
Addiction	The percent of inmates with a history of significant alcohol use (ALCOL); the percent of inmates with a history of significant drug use (DRUGS)	+, +
Previous Record	The number of previous convictions which resulted in periods of incarceration of six months or more (RECORD)	?
Marital Status	The percent of inmates who are married (MARRIED)	-
Rehabilitative Activities	The number of rehabilitative activities provided during the period and its squared value and its value interacted with CD-ALL (IPRS. IPRS2, CD-ALL* IPRS)	?, ?, ?

Table 3

The Estimated Long Run Average Cost Curve for Six FCIs Using Ordinary Least Squares

	Initial Specification		Final Specification	
Variable	Coefficient	t-ratio	Coefficient	t-ratio
Intercept	51.680161	1.2933	39.150440	4.3261***
Output				
CD-ALL	9.3807×10^{-6}	0.2418	2.155835×10^{-5}	2.2922***
LNCD-ALL	-1.681052	-0.4607	-2.449731	-3.2825***
Factor Prices				
LNCOST-L	-0.705843	-0.6858	-0.289815	-0.4740
LNCOST-C	-2.951293	-0.3530	-0.335975	-0.6408
Product Quality				
SECURE	4.126519	0.6225		
Product Quality and Service Conditions				
IPRS	-4.9511×10^{-5}	-0.4108		
IRRS2	-5.80347×10^{-4}	-0.0300		
CD-IPRS	6.424353×10^{-3}	0.3058		
AGE-I	-0.182319	-0.3024		
AGE-I2	1.730014×10^{-3}	0.1810		
RACE-I	-0.071175	-1.7084*		
BETA-IQ	-0.044039	-0.4459		
WCOLLAR	0.029246	0.7696		
LENGTH	0.162459	0.7699		
O-PERS	-0.086788	-0.8333		
O-PROP	0.028065	1.0395		
DRUGS	4.752206×10^{-4}	0.0193		
ALCOHOL	0.077491	1.3125		
RECORD	0.103382	0.3002		
MARRIED	0.016079	0.5384		
FEMALE	0.046490	1.9028*	0.018893	3.7500***
Service Conditions				
R-BAL	-4.601240	-1.1431		
R-BAL2	1.214325	0.3213		
PC-OTH	-0.029699	-0.4496		
Labor Quality				
RATIO-S	0.462797	0.3732		
AGE-S	-0.278803	-2.0947**	-0.176617	-4.8028***
ED-S	-0.036526	-0.2132		
RACE-S	0.066769	1.0285		
SEX-S	-4.0059×10^{-5}	-0.3333		

Table 3

(cont'd)

Variable	Coefficient	t-ratio	Coefficient	t-ratio
Capital Quality				
SQFPER	-0.132315	-1.1588	-0.033156	-2.2100**
SINGLE	-2.931725	-0.7770	-1.573329	-5.2792**
SANPER	2.843853	0.9086	2.124388	3.1417**
R^2(F-ratio)	0.8704	(4.61)	0.8126	(24.08)
N	60		60	

*Indicates that the coefficient was significant at the .10 level, two tail test.
**Indicates that the coefficient was significant at the .05 level, two tail test.
***Indicates that the coefficient was significant at the .01 level, two tail test.

Table 4

Results of Estimating the Frontier Average Cost Function

Variable	Coefficient	"t-ratio"
Intercept	29.6038	4.09
Output		
CD-ALL	1.615E-5	-3.23
LNCD-ALL	-2.1480	2.02
Factor Prices		
(LNCOST-C)-(LNCOST-L)	0.4305	2.56
Production Quality and Service Conditions		
FEMALE	0.0123	3.11
Labor Quality		
AGE-S	-0.1244	-5.16
Capital Quality		
SQFTPER	-0.0196	-1.39
SINGLE	-1.2488	-4.73
SANPER	1.2230	2.50
Estimated Variances		Standard Error
$\hat{\sigma}^2_{\psi}$ (the entire disturbance)	0.0204	0.0067
$\hat{\sigma}^2_{v}$ (the normal portion)	0.0066	0.0028
$\hat{\sigma}^2_{u}$ (the half-normal portion)	0.0138	0.0084

N 60

Table 5

Estimated Inefficiency ($\hat{u}$) by Prison and Quarter

		Prison 1	2	3	4	5	6
	1	.08	.10	.06	.05	.08	.09
	2	.08	.15	.09	.21	.07	.30
Q	3	.06	.11	.08	.04	.06	.07
U	4	.09	.02	.12	.10	.09	.07
A	5	.08	.09	.07	.07	.10	.05
R	6	.08	.09	.09	.10	.07	.07
T	7	.05	.06	.10	.08	.09	.06
E	8	.06	.06	.10	.08	.06	.08
R	9	.08	.12	.03	.08	.06	.04
	10	.19	.24	.13	.07	.19	.19
	Mean Inefficiency	.085	.104	.087	.088	.087	.102

THE OSPREY COMPANY

PERFORMANCE MEASURES FOR BUDGET JUSTIFICATIONS: DEVELOPING A SELECTION STRATEGY

by

Gloria A. Grizzle

Working Paper 83-6

August 1983

submitted to Public Productivity Review

Acknowledgement: The author would like to thank Ann G. Jones and Camille F. Rogers for their assistance in this research. This research was supported in part by Grant Number 80-IJ-CX-0033 from the National Institute of Justice, U.S. Department of Justice. Points of view or opinions stated in this document are those of the author and do not necessarily represent the official position or policies of the U.S. Department of Justice.

Performance Measures for Budget Justifications: Developing a Selection Strategy

This paper explores development of a tool that agency administrators can use for deciding which performance measures they should use. While it focuses upon providing information for budget preparation, the strategies discussed are equally applicable to selecting performance measures for other administrative activities, such as monitoring program implementation. The first two sections review why performance information is important for budget justification and the extent to which different types of performance measurements appear in budget documents. The next two sections describe how a measurement selection tool was developed and summarize the results of the experience using three variations of the tool. A concluding section analyzes strengths and weaknesses of the tool and suggests alternative strategies for choosing performance measures, using the assessment tool as a guide.

The Importance of Performance Data to Budget Reform

Budget reform proponents believe that the type of information presented in proposed budgets affects budget outcomes. For example, Schick notes that the two most important aspects of budgetary technique are "the data used for making program and financial decisions and the form in which the data are classified."[1] Performance budgets, program budgets, Planning Programming Budgeting Systems, Management by Objectives, and Zero Base Budgeting are all budget reforms that require information about agency or program performance. These reforms cannot be expected to produce the results their proponents anticipate when performance data are lacking.

The Extent to Which Performance Data Currently Appear in Budget Documents

A sampling of budget documents from jurisdictions that have "implemented" one or more of these budget reforms will convince the reader that changing the budget's format is more often accomplished than changing the information presented in the budget. Many of these documents still rely heavily upon workload measures as evidence of agency or program performance to the near exclusion of information about efficiency, cost-effectiveness, equity, and quality of service delivery. A survey of 88 cities, for example, revealed that 74% used a performance budget format but only 31% used efficiency information when making spending choices.[2] As a further case in point, consider Lauth's findings of the status of performance measurement in Georgia after ten years of Zero Base Budgeting:

> A perusal of the evaluation measures actually submitted by agencies indicates that with few exceptions they are workload or output measures. . . . Far less frequently . . . do the measures provide evidence about the degree to which a program economically manages the workload associated with meeting its objectives by identifying anything resembling per unit cost of production, activity or output. Rarely, if ever, are the measures indicators of program effectiveness in the sense of identifying the impact of a program on the target population or clientele.[3]

One reason[4] for relying upon workload statistics rather than efficiency or cost-effectiveness measures is that someone is already regularly collecting workload data but no one is regularly collecting service quality, efficiency, equity, and cost-effectiveness data. Collecting these other types of performance data can be expensive. The potential cost of collecting and reporting performance information suggests that agencies must be selective when collecting performance data, choosing only those measures that are worth their cost.

On what basis should agencies choose which performance measures to include in their budget justifications? Some budget offices have included in their budget preparation guidelines specific criteria that agencies should use when evaluating the suitability of potential performance measures. The State of Wisconsin and the City of Tallahassee, Florida, provide two examples. Wisconsin's guidelines for its program budget in 1971-73 stipulated that performance measures should be output-oriented, relevant to program objectives, capable of meaningful quantification, thoroughly defined, simple but informative, available on a continuing basis, and should test the validity of objectives and recognize different levels of performance.[5] Tallahassee's guidelines for its 1979 productivity budget suggested that potential measures be evaluated in terms of validity, utility, timeliness, acceptability, simplicity, and availability.[6] Most budget offices, however, provide no specific selection criteria.

A Tool for Choosing Performance Measures

The objective of this research was to develop a tool that agencies could use to screen potential performance measures systematically in order to choose measures worth including in their budget request justifications. The tool developed should be capable of discriminating among measures in terms of specific criteria. It should also be fairly easy and quick for agency personel to use.

The first step in developing this tool was to identify the criteria against which performance measures should be evaluated. To determine whether a consensus existed about the appropriate criteria to use when choosing "good" measures, we reviewed 24 books and articles on performance measurement.[7] Table 1 lists those criteria cited in more than one

Table 1

Most Frequently Cited Criteria for Choosing Good Measures, Based on Literature Survey

Criterion	Number of Times Cited
Validity	15
Clarity	14
Reliability	13
Relevance to objectives, decisions	11
Accuracy	10
Sensitivity	8
Cost	7
Ease of obtaining data	7
Precision	6
Controllability	6
Timeliness	6
Completeness	5
Uniqueness	5
Comparability	5
Consistency	3
Credibility	3
Usefulness	3
Ability to monitor quality of data	2
Privacy	2
Flexibility	2
Representativeness	2
Importance	2

article. Validity was the most frequently cited criterion, occurring in 15 of 24 articles. Clarity and reliability were also cited in over half the articles.

Next, the most freqently cited criteria were classified into four categories: technical adequacy, practicality, and two utility categories. The criterion "precision," although cited six times, was believed to be a function of sensitivity and reliability and was therefore not included as a separate criterion. Two criteria, completeness and uniqueness, were considered components of the critierion "validity." Except for these modifications, all criteria cited five or more times were included in the assessment instrument.

Table 2 lists these criteria and includes a question or two that should be answered in order to evaluate each potential measure against a specific criterion. Criteria used to evaluate technical adequacy permit assessing potential measures in terms of how valid, reliable, and accurate the measurements are likely to be. Criteria for practicality address concerns about the cost and ease of obtaining data.

Utility criteria need to be divided into two categories. One category can be applied without knowing who will use the measure being assessed and the purpose for which it will be used. This category permits assessing the extent to which the measures are clear, sensitive, and comparable. A second category of utility criteria cannot readily be used unless one first knows something about the user and the purpose to which the measurements will be put. This category assesses a measure in terms of its relevance to the decision to be made; whether the information can be provided before the decision is made; and whether the aspect of performance indicated by the measure is susceptible to control by the program, agency, or government whose performance is being measured.

Table 2
Criteria Included in the Assessment Tool

I. Technical Adequacy

A. Valid
Does the measure logically represent the concept or construct to be measured?

1. Complete
Does the measure cover the entire concept or construct?

2. Unique
Does the measure represent some concept or construct not covered by any other measure in this set?

B. Reliable
If a measurement is repeated, will the results be identical? Are there fluctuations in the characteristic to be measured, changes in transient personal or situational factors, or inconsistencies in the measurement procedure that cause variation in the measurement obtained?

C. Accurate
Is the measurement free of systematic error or bias?

II. Practicality

A. Cost
How much will data collection or analysis cost?

B. Ease of data collection
What is the anticipated ease or difficulty of obtaining data needed to make the measurement?

III. Utility - User Independent

A. Comparable
Can this measure be used to compare different programs with each other?

B. Sensitive
Is the discriminating power of the measurement procedure sufficient to capture the variation that occurs in the object, event, or situation being measured?

C. Clear
Can the meaning of the measure be understood?

IV. Utility - User Dependent

A. Relevant to Decision
Does the measure provide information needed to make a decision about the performance of a program or agency?

B. Timely
Are changes in the objects, events, or situations being measured reflected quickly enough in the measurements to be available before the decision must be made?

C. Controllable
To what extent can the user of the measure affect the measurements, providing resources are made available?

Three versions of a performance measures assessment instrument were developed. Table 3 summarizes each version's major characteristics. In version A, a three-point scale was developed for each criterion, borrowing heavily from the assessment tool reported by Blair.[8] For each criterion, three categories were defined. For example, a measure for the criterion "accuracy" would be judged to fall within one of these three categories:

High = Measurement has little or no systematic error.

Medium = Size of systematic error is known and constant across time periods.

Low = Systematic error is known to be present. Its size is either large or unknown, and constancy across time periods is undetermined.

Using this scale, a person must judge the degree to which a proposed measure meets each criterion as being either high (scored 2 points), medium (1 point), or low (0 points). A total overall score for each measure could therefore range from 0 (if the rater judges the measure as being low on all 12 criteria) to 24 (if the rater judges the measure as being high on all 12 criteria). The resulting overall scores could then be used to rank a list of performance measures in terms of overall adequacy.

An advantage of version A is that the categories defined for each scale encourage a consistent thought process across different raters and across different measures. Nevertheless, this version, as is the case for the other two versions, is subjective. Depending upon their knowledge of the measure being assessed, two people might assign a different score to the same measure. Two possible problems — imperfect rater knowledge and lack of rater diligence — could limit the usefulness of all three assessment instrument versions.

Table 3

Performance Measures Assessment Instrument:
Characteristics of Three Versions

Version of Assessment Instrument	Characteristics of Each Version		
	Criteria Used	Method for Criterion-Specific Ratings	Total Score for Each Measure
A	Instrument stipulates criteria used	Rater must apply 3-point scale for each criterion based on defined categories supplied with instrument	Overall ranking is by summing scores on individual criteria
B	Instrument stipulates criteria used	Rater must judge each measure as being satisfactory or unsatisfactory on each criterion	Wholistic rating made after rating for each criterion
C	Rater supplies own criteria to substantiate his wholistic rating	None	Wholistic rating is first step in the assessment process

Version B is similar to version A in that the rater first assesses a measure in terms of the same 12 criteria and afterwards gives the measure an overall numerical score. Version B differs from version A in two respects. Instead of using a three-point scale, the Version B user judges each measure as being either satisfactory or unsatisfactory on each criterion. Definition of the terms "satisfactory" and "unsatisfactory" is left to the rater. A second difference is how the rater determines the overall score. After considering a measure's adequacy according to each criterion, the rater assigns the measure a rating from 0 to 10. Version B therefore allows the rater to assign an overall rating that reflects his opinion of the relative importance of the various criteria. It also allows the rater to base his rating on other factors in addition to the criteria stipulated in the assessment instrument.

Version C reverses the steps in the rating process. The rater first considers a measure and assigns an overall numerical rating from 0 to 10. The rater then lists his reasons for the rating assigned. The instrument does not stipulate the criteria that the rater must use.

Experience with the Assessment Instrument

Several groups of people have used one or more versions of this assessment tool. Table 4 summarizes their experience. Six students in a graduate program evaluation class used version A. Each student generated his own measures as a part of an evaluation design for a public-sector program. After about four hours of discussion about performance measures and measurement criteria, each student used version A to rank the performance measures that he developed. This ranking was done during whatever time the student chose the week following the discussions. Total

Table 4

Summary of Experience When Using
Performance Measures Assessment Measures

Characteristic	Version of Assessment Instrument A	B	C
Number of measures rated			
Public administration students	36-92	10	10
Prison staff		12	12
Number of raters			
Public administration students	6	8	11
Prison staff		10	10
Median minutes required to rate each measure			
Public administration students	3.0	2.0	1.5
Prison staff		1.7	1.4
Satisfaction with method used			
Public administration students	Slightly satisfied	Slightly satisfied	Slightly satisfied
Prison staff		Neutral	Slightly satisfied
Distribution of ratings*			
Public administration students	5,13,16,19,23	0,3.5,6,8,10	0,3,6,7,10
Prison staff		0,5,7,8,10	0,4,5,7,10

* Statistics are listed in the following order: low score, first quartile, median, third quartile, high score. Possible range for version A is 0-24; for versions B and C, 0-10.

rating time ranged from 2½ to 4 hours, averaging 3 minutes for assessing each measure.

Two groups of people used version B, 8 being students in another graduate program evaluation course and 12 being staff members in a Federal prison. The majority of the students were fulltime employees of a state government. Both groups used a list of proposed measures the researcher furnished them. The students individually rated 10 performance measures for a probation program for which they were developing an evaluation design as a class project. As was the case for verision A, about 4 hours of discussion about performance measurement and measurement criteria preceded the rating session.

The prison staff used version B to rate 12 performance measures for a prison. Each staff member was approached individually, the purpose of the instrument explained to him, and the definition of each criterion given him in writing. A researcher was present and available to answer questions when the staff person assessed the measures.

Two groups of people also used version C. Eleven students who used version C include the 8 who used version B. They applied version C to the same set of probation performance measures used in version B. They used version C a week before they saw the version B instrument. At the same Federal prison, a different group of staff members used version C to rate the same prison measures rated by the other group with version B. Prison staff also used version C individually in the presence of a researcher after listening to an oral explanation.

As Table 4 indicates, there is not much difference among the three versions in either the time required or the level of satisfaction with the instrument the users reported. Two additional factors that need to be

explored in more detail, however, are the instrument's discriminating power and its subjectivity. If people assign most measures similar scores, then the instrument is not a useful tool for choosing adequate measures. The histograms in Figure 1 distribute the scores for the five trials summarized in Table 4. Figure 2 displays central tendency and dispersion statistics for these histograms. The number in the middle of each box in Figure 2 is the median score for the trial. The numbers on the left and right sides of each box are the first and third quartiles, respectively. The numbers at the left and right ends of each line are the low and high scores, respectively. Thus, 50% of the scores are within the range delineated by the box, and 25% of the scores are higher and another 25% are lower than this range. Both the histograms and summary statistics show that there is enough spread in the scores to discriminate among potential performance measures.

As noted previously, assigning scores is a subjective act. Saaty states that objectivity means shared subjectivity in interpreting experience.9 Accepting this definition of objectivity permits measuring an instrument's objectivity by the extent to which different people's scores for a given measure agree. Figures 3 and 4 display the dispersion and central tendency statistics for the trials that used versions B and C of the instrument. The anchor points for the scores summarized in these figures are 0 and 10. For both versions the lack of agreement in scoring measures is substantial, suggesting that assigning scores is indeed subjective.

Conclusion

Why do people differ in the scores they assign a given measure? There are four major sources of disagreement: (1) People may consider different attributes of a measure when scoring it. (2) People may differ in terms of

Figure 1

HISTOGRAMS OF SCORES FROM TRIALS USING
THE PERFORMANCE MEASURES ASSESSMENT INSTRUMENT

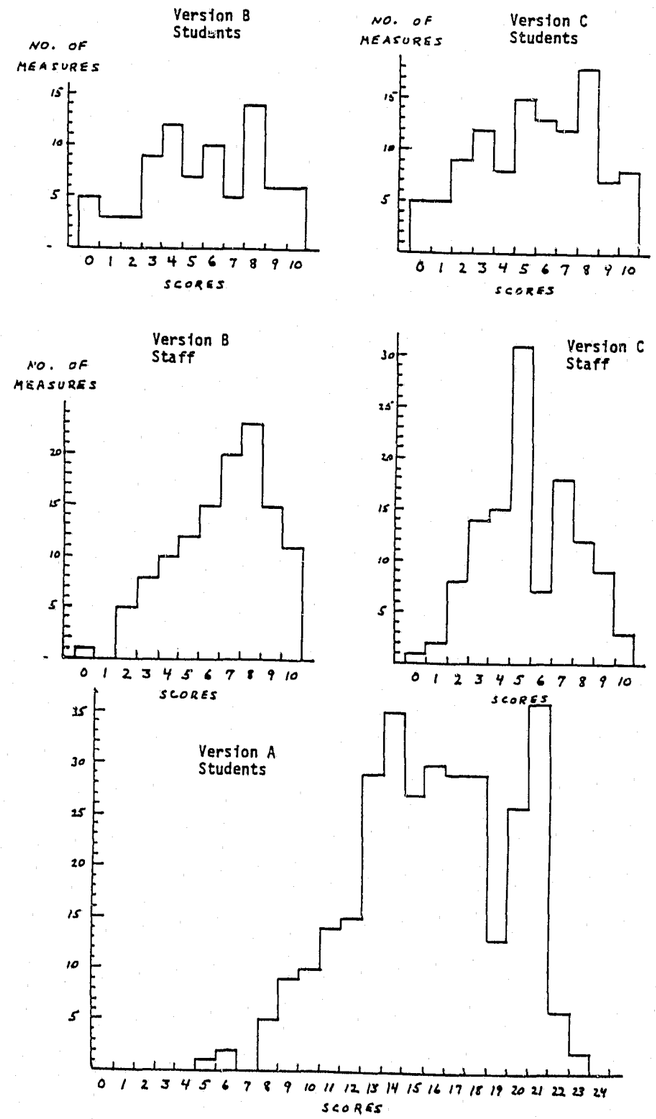
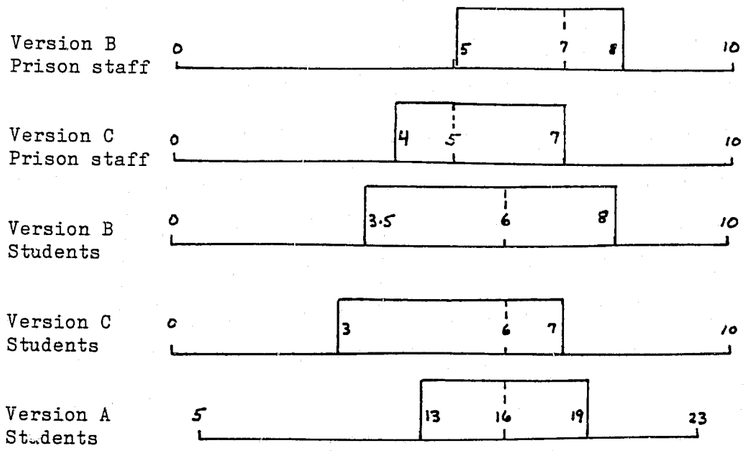


Figure 2

Distribution of Scores Around the Median from Trials Using the Performance Measures Assessment Instrument



How to read this diagram: The number in the middle of each box is the median. The numbers on the left and right sides of each box are the first and third quartiles, respectively. The numbers at the left and right ends of each line are the low and high scores, respectively. Thus, 50% of the scores are within the range delineated by the box, and 25% of scores are higher and another 25% are lower than this range.

Figure 3

DISTRIBUTION OF SCORES AROUND THE MEDIAN, BY MEASURE, FROM STUDENTS USING THE PERFORMANCE MEASURES ASSESSMENT INSTRUMENT

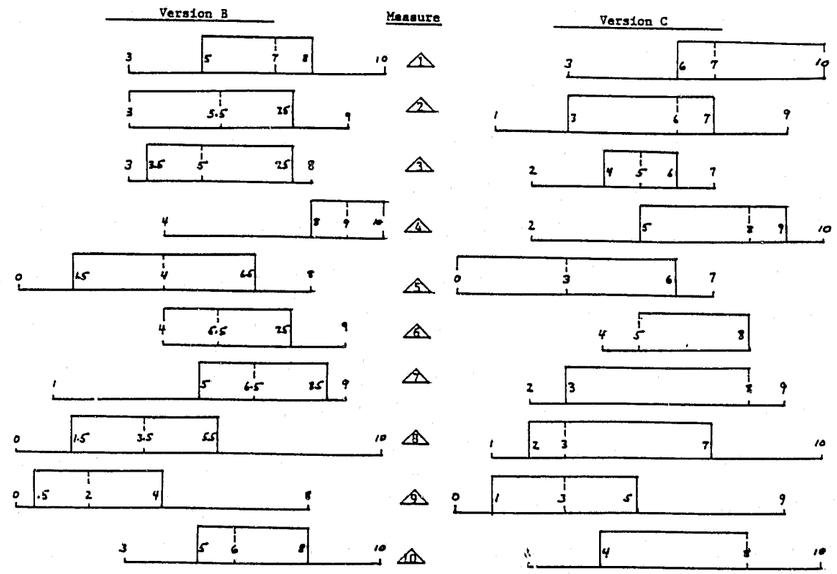
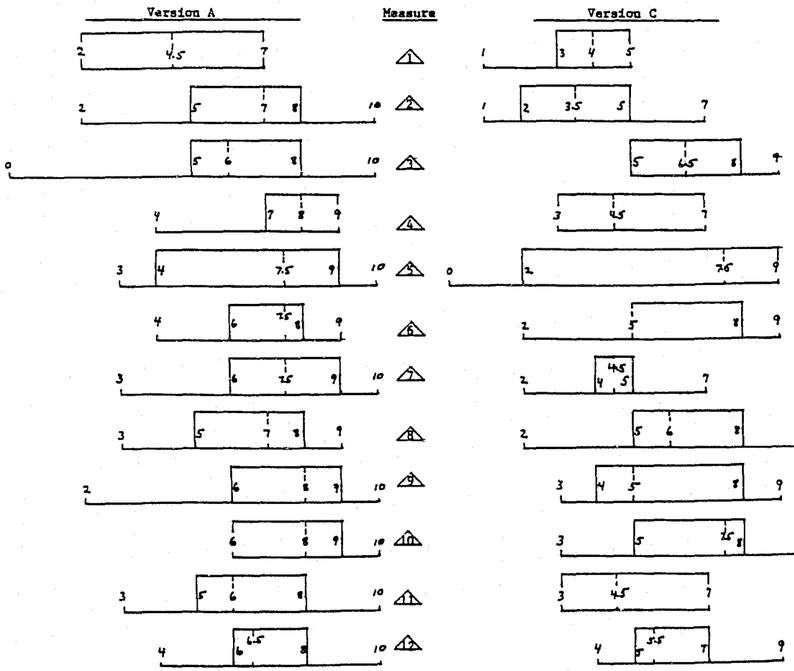


Figure 4

DISTRIBUTION OF SCORES AROUND THE MEDIAN, BY MEASURE, FROM PRISON STAFF USING THE PERFORMANCE MEASURES ASSESSMENT INSTRUMENT



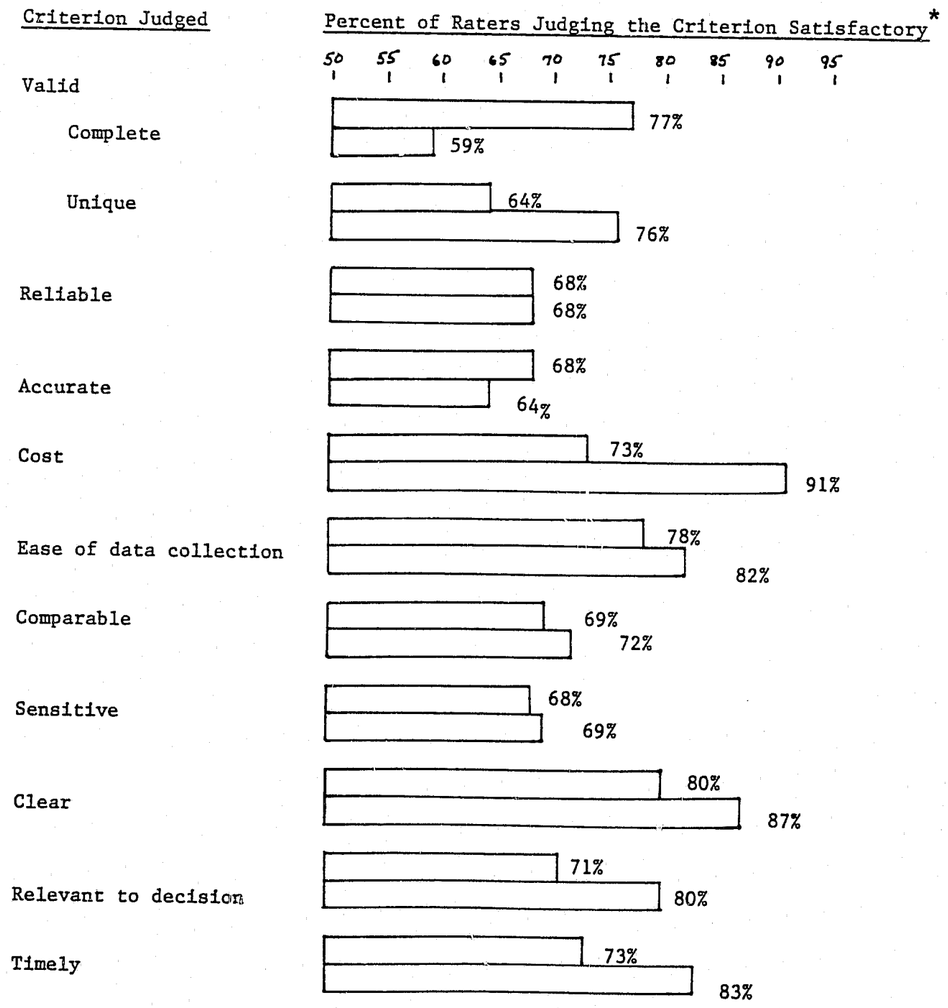
the relative importance they ascribe to the attributes they consider. (3) People may have imperfect knowledge about a measure's attributes. (4) People may inconsistently apply the criteria by which they assess a measure's attributes. These sources of subjectivity suggest that one might follow several strategies in order to reduce the tool's subjectivity when using it to choose performance measures.

Using version A should eliminate the first two sources of disagreement because it prescribes the attributes (i.e., the criteria) upon which its users assess measures and also, through the three-point scale for each criterion, prescribes that all attributes receive equal weight. Another approach to eliminating the first two sources of disagreement is to have the same person or team rate all the measures being considered. The amount of disagreement among raters using versions B and C suggests that it would be inappropriate when using these versions to have one person score part of the measures, someone else to score the rest, then combine both sets and choose the measures with the highest scores. When the first two sources of disagreement are not controlled for when scoring measures, measures should not be compared with each other on the basis of the scores unless the same person or group scored them all.

Using version A does not eliminate the third and fourth sources of disagreement. These two sources may be a problem with all three versions. Figure 5 shows the percentage of agreement among raters for the two groups that used version B, which required judging a measure to be either satisfactory or unsatisfactory on each criterion. The greatest possible disagreement is for half the group to rate a measure as being satisfactory on a given criterion and for the other half of the group to rate it as being unsatisfactory on the same criterion. One might expect that the prison

Figure 5

RATING PERFORMANCE MEASURES USING VERSION B -
EXTENT OF AGREEMENT BY CRITERION



*The top and bottom bar for each criterion shows agreement from student trial and prison staff trial, respectively.

staff would have more information about prison measures than students would have about probation measures and that the prison staff would therefore agree with each other more. The prison staff did show greater agreement on 8 criteria, but less agreement on 2 criteria.

When people using version A were unsure of how to score a measure in terms of the three-point scale, they tended to assign the middle point in the scale. The effect of this tendency is that lack of information about the measures' attributes results in measures receiving similar scores. It may be possible for some agencies to have their staff specialize when assessing measures in order to make more informed judgments. One basis for specialization might be to have one person assess the technical adequacy of the proposed measures, another person assess their practicality, a third assess their utility, and a fourth give the measures an overall score based upon the ratings of the other three.

The last source of disagreement, inconsistent application of the criteria, may occur when a person does not understand how to use the instrument or is not diligent when using it. Appropriate explanation and training should solve the first problem. When agency staff understand the use to which their assessments will be put, they may feel they have enough stake in the outcome to undertake the task with reasonable diligence.

We have already noted that people may vary in the relative importance they ascribe to different measurement criteria. Table 5 summarizes the opinions of a convenience sample of two groups — staff in a state planning and budgeting office and staff in a Federal prison. This small sampling can in no way be generalized to broader groups of people, but it does show that people may differ in the relative importance they accord the criteria stipulated in versions A and B of the assessment tool.

Table 5

Opinions about the Relative Importance of Attributes of Performance Measures

Attribute	Percentage Distribution of Responses*			
	Essential	Highly Desirable	Nice but Optional	Not Important
Complete	17/50	33/33	50/8	0/8
Unique	0/8	50/33	50/58	
Reliable	83/75	17/25		
Accurate	100/67	0/33		
Cost	0/8	100/33	0/42	0/17
Ease of data collection	0/8	100/25	0/58	0/8
Comparable	0/17	83/67	17/17	
Sensitive	17/33	83/58	0/8	
Clear	83/67	17/33		
Relevant to decision	67/50	33/33	0/17	
Timely	33/33	50/67		

* The percentage to the left of the oblique represents the responses of a convenience sample of 6 planning and budgeting staff. The percentage to the right of the oblique represents the responses of a convenience sample of 12 prison staff.

When such is the case, they are unlikely to find implementing version A (which accords equal weight to each criterion) a satisfactory approach.

When agencies believe some criteria are more important than others, they can modify the assessment tool to economize upon the assessment task. By using the most important criteria as a screening device, the total number of measures can thereby be reduced to a subset that merits further assessment. Versions A and B have in fact been adapted in this fashion. In one instance, version A was adapted to screen about 1100 potential measures for corrections programs.10 In the first step, measures that scored low on the validity criteria (completeness and uniqueness) were discarded. In the second step, the remaining measures were further assessed in terms of reliability, accuracy, comparability, sensitivity, and clarity. In another instance, version B was adapted to rate about 500 potential measures being considered for a state's social programs (education, health, social services).11 In this instance, a two-person team assessed each measure, again using a two-step procedure. In the first step, the team selected measures on the basis of completeness and clarity. They next took those measures rated satisfactory in terms of these two criteria and rated them in terms of accuracy, uniqueness, and cost of data collection.

Potential measures need to be assessed by people who understand the context in which performance measures submitted in budget justifications will be used. Questions of practicality and relevance to resource allocation decisions may need to be weighed more heavily than would be the case for research applications. The tool described in this paper gives one a systematic way of thinking about factors that render a potential measure adequate or inadequate for a given situation. As the applications mentioned demonstrate, the tool can be adapted to develop an asessment strategy

CONTINUED

2 OF 3

appropriate to an agency's concerns, staff skills, and resources available for data collection. Applied systematically, such an assessment instrument can identify from a list of potential measures those worth including in agency budget justifications. As such, it can be a helpful tool in facilitating budget reform implementation.

Footnotes

1. Allen Schick, "The Road from ZBB," Public Administration Review, 38:2 (March/April, 1978), p. 178.

2. Lewis Friedman, "Performance Budgeting in American Cities," Public Productivity Review, 3:4 (Spring/Summer, 1979), 50-51.

3. Thomas P. Lauth, "Performance Evaluation in the Georgia Budgetary Process" (Paper presented at the American Society for Public Administration National Conference, April, 1981), p. 7.

4. Others have ably documented the many political, organizational, economic and behavioral problems that can also hinder collecting, and reporting performance information for budget review. See, for example, "Creative Budgeting in New York City: An Interview with Former Budget Director Frederick O'R. Hays" (Washington, D.C.: The Urban Institute 1971); Merlin M. Hackbart and James R. Ramsey, "Budgeting: Inducements and Impediments to Innovations," State Government, 52:2 (Spring 1979), 65-69; Frederick O'R. Hayes, "The Budget and Its Problems," Urban Affairs Papers, 2:2 (Spring 1980), pp. 7-18; Thomas P. Lauth, "Zero-Base Budgeting in Georgia State Government: Myth and Reality," Public Administration Review, 38:5 (September/October 1978), 420-430; Perry Moore, "Zero-Base Budgeting in American Cities," Public Administration Review, 40:3 (May/June 1980), 253-258; A. Premchand, "Government Budget Reforms: Agenda for the 1980s," Public Budgeting and Finance, 1:3 (Autumn 1981), 16-24; A. Premchand, "Government Budget Reforms: An Overview," Public Budgeting and Finance, 1:2 (Summer 1981), 74-85; Richard Rose, "Implementation and Evaporation: The Record of MBO," Public Administration Review, 37:1 (January/February 1977), 64-71; Allen Schick, "A Death in the Bureaucracy: The Demise of Federal PPB," Public Administration Review, 33:2 (March/April 1973), 146-156; Allen Schick, "The Road from ZBB," Public Administration Review, 38:2 (March/April 1978), 177-180; Allen Schick, "The Road to PPB: The Stages of Budget Reform," Public Administration Review, 26:4 (December 1966), 243-258; Elmer Staats, "The Continuing Need for Budget Reform" (address to the American Association for Budget and Program Analysis, 1980); Jeffrey D. Straussman, A Typology of Budgtary Environments: Notes on the Prospects for Reform," Administration and Society, 11:2 (August 1979), 216-226; Paul T. Veillette, "A Public Accounting: Reflections on State Budgeting," Public Budgeting and Finance, 1:3 (Autumn 1981), 62-68; Aaron Wildavsky, "The Political Economy of Efficiency: Cost-Benefit Anlaysis, Systems Analysis, and Program Budgeting," Public Administration Review, 26:4 (December 1966), 292-310; and Aaron Wildavsky and Arthur Hammond, "Comprehensive Versus Incremental Budgeting in the Department of Agriculture," Administrative Science Quarterly, 10:3 (December 1965), 321-346.

5. State of Wisconsin, "Manual on Program Budget Preparation" (S. Kenneth Howard and Gloria Grizzle, eds. Whatever Happened to State Budgeting? Lexington, Ky.: Council of State Governments, 1972), pp. 255-256.

6. City of Tallahassee, Florida, "Productivity Measurement Worksheet," (1979).

7. See William Ascher, Forecasting: An Appraisal for Policy-Makers and Planners (Baltimore: Johns Hopkins University Press, 1978); Louis H. Blair, et al., Monitoring the Impacts of Prison and Parole Services: An Initial Examination (Washington, D.C.: The Urban Institute, 1977); Milton M. Chen, J. W. Bush, and Donald L. Patrick, "Social Indicators for Health Planning and Policy Analysis," Policy Sciences 6:1 (March 1975), 71-89; Thomas J. Cook, "Performance Measures for the Courts System" (Triangle Park, N.C.: Research Triangle Institute, 1978), grant application submitted to Law Enforcement Assistance Administration; George S. Day and Burton A. Weitz, "Comparative Urban Social Indicators: Problems and Prospects," Policy Sciences 8:4 (December 1977), 423-435; John J. Dinkel and Joyce E. Erickson, "Multiple Objective in Environmental Protection Programs," Policy Sciences 9:1 (February 1978), 87-96; Allan R. Drebin, "Criteria for Performance Measurement in State and Local Government," Governmental Finance, 9:4 (December 1980), 3-7; Gloria A. Grizzle, et al., "Performance Measurement Theory for Corrections" (Raleigh, N.C.: The Osprey Company, 1978), grant application submitted to the Law Enforcement Assistance Adminstration, 1978; Owen P. Hall, Jr., "A Policy Model Appraisal Paradigm," Policy Sciences, 6:2 (June 1975), 185-195; Harry P. Hatry, "Performance Measurement Principles and Techniques: An Overview for Local Government," Public Productivity Review, 4:4 (December 1980), 312-339; Peter J. Hunt, Program Evaluation Manual (Madeira Beach, Fla.: Personnel Research and Training Institute, 1978); E. Gerald Hurst, Jr., "Attributes of Performance Measures," Public Productivity Review, 4:1 (March 1980), 43-49; Joan Jacoby, "Theory of Performance Measurement for Prosecution and Public Defense" (Washington, D.C.: Bureau of Social Science Research, Inc., 1978), grant application submitted to Law Enforcement Assistance Administration; Abraham Kaplan, The Conduct of Inquiry: Methodology for Behavioral Science (San Francisco: Chandler, 1964); Helmut Klages, "Assessment of an Attempt at a System of Social Indicators," Policy Sciences 4 (1973), 249-261; Delbert C. Miller, Handbook of Research Design and Social Measurement, 2nd ed. (New York: David McKay Co., 1970); Jum C. Nunnally and Robert L. Durham, "Validity, Reliability, and Special Problems of Measurement in Evaluation Research," in Elmer Struening and Marcia Guttentag (eds.), Handbook of Evaluation Research, Vol. 1 (Beverly Hills: Sage 1975), 289-352; Dale K. Sechret, "The Development and Implementation of Standards for Correctional Systems" (College Park, Md.: American Correctional Association, 1979); Claire Selltiz, et al., Research Methods in Social Relations (New York: Holt, Rinehart and Winston, 1959); Robert L. Thorndike and Elizabeth P. Hagen, Measurement and Evaluation in Psychology and Education (New York: John Wiley, 1977); Gordon P. Whitaker, and Elinor Ostrom, "Performance Measurement in the Criminal Justice System: A Police Perspective" (Chapel Hill:

University of North Carolina), grant application submitted to the Law Enforcement Assistance Administration, 1978; Marshall H. Whithed, "Toward the Development of a Set of World Criminal Justice Indicators" (Virginia Commonwealth University, unpublished manuscript, March 1969); Harold L. Wilensky, Organizational Intelligence (New York: Basic Books, 1967); and Joseph W. Wilkinson, "The Meaning of Measurements," Management Accounting, 57 (July 1975), 49-52.

8. Louis H. Blair, et al., Monitoring the Impacts of Prison and Parole Services: An Initial Examination (Washington, D.C.: The Urban Institute, 1977).

9. Thomas L. Saaty, The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation (New York: McGraw-Hill, 1980), p. 15.

10. Ann G. Jones, "The Rating of Corrections Performance Measures" (Raleigh, N.C.: The Osprey Company, 1980).

11. Gloria A. Grizzle and Karen S. Minerva, "Measuring the Cost-Effectiveness of Florida's Social Programs" (Tallahassee, Fla.: Florida State University, 1980).

THE OSPREY COMPANY

DEVELOPING PERFORMANCE-DIMENSION WEIGHTS FOR ASSESSING PUBLIC-SECTOR PROGRAMS: METHOD AND CONTEXTUAL EFFECTS

by

Gloria A. Grizzle

Working Paper 83- 4

July 1983

submitted to Organizational Behavior and Human Performance

Prepared under grant 80-IJ-CX-0033 from the National Institute of Justice, U.S. Department of Justice. Views and opinions are those of the author and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

DEVELOPING PERFORMANCE-DIMENSION WEIGHTS FOR ASSESSING PUBLIC-SECTOR PROGRAMS: METHOD AND CONTEXTUAL EFFECTS

Public-sector programs, and the policies that create them, almost always have multiple objectives. To seek out optimal policies systematically rather than intuitively, one needs to know the relative importance (or weights) of the objectives relevant to a given policy. Knowing these weights permits aggregating into a single overall measure of performance or utility a policy's effect upon several objectives. By comparing these overall scores for a set of policy alternatives, one can then select the "best" alternative in the set.

A number of methods have been devised that permit one to elicit from relevant parties in the policy-making process their judgments about the relative importance of these objectives. Hwang and Yoon (1981) classify seventeen major methods.

Some researchers have concerned themselves with the possible effects that choice of elicitation method may have upon the judgments elicited. Several studies that compare the results of using two or more methods have been reported in the literature. Among the methods compared are clinical or intuitive judgments; holistic or observer-derived methods, such as the social judgment or policy capturing technique; analytic or decomposed rating methods, such as simplified multiattribute rating technique, simple multiattribute utility procedure, indifference tradeoffs, and analytic hierarchy process; simple point allocation; the nominal group technique; and equal weights. The findings from these studies are mixed.

Some studies reported that the judgments produced by different methods are similar. Einhorn and McCoach (1977) found a similarity of results for ranking, rating, and equal weighting methods. Schoemaker and Waid (1982) reported the several methods they compared yielded significantly different weight estimates. They concluded, however, that holistic or observer-derived, indifference trade-off, analytic hierarchy, and point allocation methods all predicted about equally well on average and that equal weighting was clearly inferior to these four methods. Rohrbaugh (1981) judged the quality of judgments produced by social judgment theory versus a nominal group technique to be equal, but concluded that social judgment theory was better at developing consensus among group members.

Other researchers found low correlations between holistic and derived judgments (Pitz, Heerboth, and Sachs, 1980), superiority of a rating technique over holistic assessments (Eils and John, 1980), some improvement of rank weighting over equal weights (Stillwell, Seaver, and Edwards, 1981), superiority of linear statistical models to intuitive judgments (Dawes and Corrigan, 1974), and striking differences in weights derived from tradeoffs versus a rating method (Hobbs, 1980). Hobbs concluded that the choice of method had as much influence upon weights as choice of person.

Still other researchers have broadened their concern to include the effects of both the method used and the problem context. Edwards (1977) maintains that the weights derived from multiattribute methods depend upon the subject, what is being

assessed, and the purpose of the assessment. Recent studies suggest that the appropriate multiattribute method may depend upon the subjects chosen and the task definition (Wallsten and Budescu, 1983; Schoemaker and Waid, 1982; Hershey, Kunreuther, and Schoemaker, 1982). Billings and Marcus (1983) report that reducing the time alloted to the elicitation process causes some subjects to use information in a curvilinear rather than linear fashion and others to use the information interactively. Possible interactions of the measurement process and problem context lead Hershey, Kunreuther, and Schoemaker (1982) to advise convergent validation.

PROBLEM STATEMENT

This study examines the effect of four factors upon people's judgments about the relative importance of several dimensions for assessing the performance of public sector programs. These factors are the individual's role in relation to public sector programs, the way the judgment task is defined, the method chosen to elicit an individual's judgment, and the type of program being assessed. The research cited above leads us to suspect that not only may each of these factors affect the judgments made, but that these factors may interact with each other.

Subjects

The subjects who participated in this exercise can be categorized as follows:

Budget analysts working for state legislative appropriations committees and state governors' offices, hereafter referred to as funders (N = 100);

Students taking graduate-level public budgeting courses,

referred to as funder surrogates (N = 212);

Researchers at universities or other research institutions who are involved in public sector research (N = 72);

Students taking graduate-level program evaluation courses, referred to as researcher surrogates (N = 22);

Administrators and service providers in public-sector programs, referred to as practitioners (N = 155);

Students taking a graduate-level public management course, referred to as practitioner surrogates (N = 12);

Citizens who are neither researchers, students, nor government employees, referred to as the general public (N = 67).

Task

The task given the subjects was to determine the relative importance of six performance dimensions for assessing the performance of a public sector program. Each subject was instructed to make these judgments from his/her perspective as a budget analyst, researcher, practitioner, or private citizen. The students were asked to assume the role of funder, researcher, or practitioner and to make their judgments from this perspective.

Dimensions for assessing a program's performance included quantity of program output, quality of output, equity of service distribution, efficiency (unit cost of output), benefit (the effect of programs upon service recipients and others in society), and program cost. To determine the sensitivity of judgments to the way the dimensions are defined, program cost was defined in two ways. For 368 subjects, cost was defined as total program cost, or the amount of money spent on the program for all

purposes, such as salaries for program staff, supplies, travel and equipment. For 272 subjects, cost was defined as cost-effectiveness, or the cost per unit of benefit.

Method

A third factor tested was the method used to elicit subjects' judgments. Three methods were used - social judgment theory (for 88 subjects), simplified multiattribute rating technique (331 subjects), and analytic hierarchy process (221 subjects). With each method subjects were given an instrument and a sheet that defined each performance dimension and gave examples of performance measures that corresponded to each dimension. The instrument for each method is described below.

Social judgment theory is a holistic approach that infers the weights for each performance dimension from a subject's overall ratings (Hammond, 1976; Hammond, Rohrbaugh, Mumpower, and Adelman, 1977). The instrument for this method consisted of a series of profiles describing the performance of forty hypothetical public programs. Figure 1 shows a sample of these profiles. For each profile, the length of the bar opposite each dimension summarizes how well that program performed on that dimension. The best possible performance is scored 10 and the worst possible is scored 0. The subject reviews each profile and makes a judgment of the program's overall performance by giving the profile a rating between 0 (worst) and 20 (best).

To determine a subject's importance weights based upon these forty ratings, a regression equation is fitted to these data. The dependent variable is the overall rating, and the dimension scores are the independent variables. The coefficients of the

performance dimensions in the resulting equation are the subject's relative importance weights. These weights are then normalized so that they sum to 100 and can be directly compared with weights elicited by other methods.

Simplified multiattribute rating technique, the second method, is analtyic in that it elicits dimension weights by direct scaling instead of deriving them as described above (Edwards, 1979; Edwards, 1980). This instrument asks the subject first to select the performance dimension that is least important and to assign it a weight of 10. The subject then assigns weights to each of the other dimensions by comparing each to the least important dimension. For example, if the subject believes quantity of output is least important and quality of output is 2.5 times as important as quantity, the subject would assign quality a weight of 25. These weights are then normalized.

The analytic hierarchy process is also an analytic approach and elicits dimension weights through a series of pairwise comparisons (Saaty, 1980). Figure 2 displays the instrument used and explains how the subject judges the relative importance of each possible pair of dimensions. To determine a subject's importance weights based upon these comparisons, we set each subject's judgments into a matrix, as illustrated in Table 1. If the dimension in the row is more important than the dimension in the column, the magnitude is expressed as a whole number. If the dimension in the row is less important than the dimension in the column, the magnitude is expressed as the reciprocal of the whole number. The numbers below the diagonal are reciprocals of the numbers above the diagonal. Next, we take the geometric mean of

each row. This geometric mean represents the importance weight of the dimension in that row relative to the others. Finally, we normalize these means.

Program

The last factor tested was the public program for which subjects were asked to judge the relative importance of performance dimensions. Four programs were included: probation and parole (362 subjects), air pollution control (93 subjects), medicaid (94 subjects), and any public program (91 subjects).

Analysis

For each factor level, mean performance dimension weights were calculated, using multiple classification analysis. Analysis of variance and the F statistic were used to test for the statistical significances of differences for both main effects and interaction terms.

RESULTS

Overall, subjects judge benefit to be most important, followed by quality of output. Equity and cost are about equally important. Least important are quantity of output and efficiency. Mean weights across all 640 subjects are as follows:

Quantity	11%
Quality	20
Equity	15
Efficiency	11
Benefit	27
Cost	16
	100%

Table 2 shows the mean weights for each factor level.

Whether broken down by method, program type, subject's role, or cost definition, benefit is always judged most important. Quantity and efficiency never attain higher than fourth place. Quality is always in second or third place. The greatest variation is for equity, whose rank ranges from second to fifth, and cost, whose rank ranges from second to sixth.

In several instances the differences in ratings are large enough for the main effects to be statistically significant at the .05 significance level, based upon the F test. Looking first at the method factor, one finds significant differences for [illegible] performance dimensions - equity (alpha = .01) and cost (alpha = .003). Subjects using the analytic hierarchy process rate equity higher and cost lower than do subjects using the other two methods. For equity, the largest point spread is between the analytic hierarchy process (20%) and the simplified multiattribute rating technique (12%). A similar difference holds for cost (10% for the analytic hierarchy process compared to 19% for simplified multiattribute utility technique).

Differences by program type were significant for only the equity dimension (alpha = .02). Subjects considered equity to be least important in assessing the performance of air pollution programs (13%) and most important for mediciad programs (18%).

Differences by subject's role are large enough to be statistically significant for half the performance dimensions - equity (alpha = .003), efficiency (alpha = .03), and cost (alpha = .004). Looking first at all seven roles, one finds that practitioner surrogates judge equity more important and the general public judges it less important - 20% compared to 12%.

Researcher surrogates and practitioner surrogates judge efficiency more important and researchers judge it less important - 13% compared to 8%. Last, funders consider cost more important and practitioner surrogates consider it less important (21% compared to 10%).

Recall that the surrogates are graduate students asked to assume the role of funder, researcher, or practitioner. The mean differences between their weights and those of the real funders, researchers, and practitioners suggest that they may not have been well socialized into their roles when they took part in this exercise. For all three performance dimensions for which differences in ratings are large enough to be statistically significant at the .05 alpha level, surrogates are at either the low or high end of the range. If we choose to ignore the surrogates and focus only upon the other four subject role levels, we find the following differences:

Researchers judge equity more important (19%) and the general public judges it less important (12%).

Funders judge efficiency more important (12%) and researchers judge it less important (8%).

Funders judge cost more important (21%) and both practitioners and the general public judge it less important (16%).

Definition of the cost dimension is the last factor whose main effect was tested. Although there was a 5 percentage difference between judgments for total cost and cost-effectiveness, this difference was not statistically significant at the .05 alpha level.

Two-way interactions between the factors were generally not statistically significant. The single exception was for the efficiency dimension, were the interaction between elicitation method and program type was significant (alpha = .02).

DISCUSSION

These findings suggest that equity is the performance dimension whose weight is most likely to be affected by one's choice of elicitation method, subject, and program type. The tradeoff seems to be between equity and cost or efficiency.

There is no obvious reason why the analytic hierarchy process would influence subjects to weight equity higher and cost lower. It is in fact the method one would expect to yield the least extreme judgments. Social judgment theory allows the subject to ignore some dimensions if he/she so chooses and to thereby give those dimensions ignored a weight of zero. Simplified multiattribute rating technique permits the subject to rate the dimension judged most important as many times greater than the least important as he/she chooses. Analytic hierarchy process, on the other hand, requires that all dimensions be given some weight and that no dimension be weighted more than nine times as important as any other.

Differences in mean weights classified by subject's role are more expected. Only funders rate cost as the second most important dimension. This importance given to the cost dimension conforms to our expectation of how budget analysts behave. It is not surprising that researchers weighted equity higher than did funders. The research subjects were predominantly from the disciplines of sociology and political science, both disciplines

concerned with the equity issue (Bodily, 1978; Coulter, 1980; Jones, 1981; Lineberry and Welch, 1974; Ostrom, Parks, Percy, and Whitaker, 1979; Wilenski, 1980-81). We are at a loss to explain why the general public gave so little importance to equity.

Program type differences on the equity dimension are not large. What differences do exist, however, seem plausible. Equity was seen to be most important for a health care service which should be available to people applying for care who meet the eligibility requirements. Equity was least important for air pollution control, which benefits all who breathe, regardless of whether they apply for the service.

In conclusion, we found that in most instances the elicitation method, subject's role, program type, and definition of cost did not affect judgments about the relative importance of performance dimensions enough to be statistically significant at the .05 alpha level. Further, interactions between these factors were generally not statistically significant. Exceptions were for equity, where elicitation method, program type, and subject's role had an effect; cost, affected by elicitation method and subject's role; and efficiency, affected by subject's role. The single statistically significant interaction term was for efficiency, where the interaction of method and program affected subject's judgments.

Figure 1

Profiles of Hypothetical Public Programs

Program 1	Quantity of output	XXXXXXXXXXXXXXXXXX
	Quality of output	XXXXXX
	Equity	XXXXXXXXXXXX
	Efficiency	XXXXXXXXXXXXXXXXXXXX
	Benefit	XXXXXXXXXXXXXXX
	Total cost	XXX
		1 2 3 4 5 6 7 8 9 10
Program 2	Quantity of output	XXXXXXXXXXXXXXXXXX
	Quality of output	XXXXXXXXXXXXXXX
	Equity	XXX
	Efficiency	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
	Benefit	XXXXXXXXXXXXXXXXXXXXXXXXX
	Total cost	XXXXXXXXXXXXXXX
		1 2 3 4 5 6 7 8 9 10
Program 3	Quantity of output	XXXXXX
	Quality of output	XXXXXXXXXXXXXXXXXXXX
	Equity	XXXXXXXXXXXXXXXXXXXXXXXXX
	Efficiency	XXXXXXXXX
	Benefit	XXXXXXXXXXXX
	Total cost	XXXXXXXXXXXXXXXXXX
		1 2 3 4 5 6 7 8 9 10

Figure 2

Instrument Used for Analytic Hierarchy Process

INSTRUCTIONS

Assume that your task is to determine the performance of a probation and/or parole agency. Use the matrix below to compare the importance of six performance dimensions as indicators of agency performance. Definitions of these dimensions appear on the lefthand side of this sheet.

Each row in this matrix compares two performance dimensions. For each row, check the column that most closely reflects your opinion of the importance of the performance dimension in the lefthand column compared with the performance dimension in the righthand column. For example, in the first row, a check in column +5 means that you believe quantity of output is strongly more important than quality of output. A check in column -3 means that quantity is moderately less important than quality. A check in column 1 means that the two performance dimensions are of equal importance as indicators of agency performance.

Absolutely more important	Very Strongly more important	Strongly more important	Moderately more important	Equally important	Moderately less important	Strongly less important	Very Strongly less important	Absolutely less important	
+9	+7	+5	+3	1	-3	-5	-7	-9	
Quantity	—	—	—	—	—	—	—	—	—	Quality
Quantity	—	—	—	—	—	—	—	—	—	Equity
Quantity	—	—	—	—	—	—	—	—	—	Efficiency
Quantity	—	—	—	—	—	—	—	—	—	Benefit
Quantity	—	—	—	—	—	—	—	—	—	Cost-effectiveness
Quality	—	—	—	—	—	—	—	—	—	Equity
Quality	—	—	—	—	—	—	—	—	—	Efficiency
Quality	—	—	—	—	—	—	—	—	—	Benefit
Quality	—	—	—	—	—	—	—	—	—	Cost-effectiveness
Equity	—	—	—	—	—	—	—	—	—	Efficiency
Equity	—	—	—	—	—	—	—	—	—	Benefit
Equity	—	—	—	—	—	—	—	—	—	Cost-effectiveness
Efficiency	—	—	—	—	—	—	—	—	—	Benefit
Efficiency	—	—	—	—	—	—	—	—	—	Cost-effectiveness
Benefit	—	—	—	—	—	—	—	—	—	Cost-effectiveness

Table 1

Illustrative Matrix Constructed from a Subject's Responses

Using the Analytic Hierarchy Process Instrument

Performance Dimension	Quantity	Quality	Equity	Efficiency	Benefit	Cost
Quantity	1	3	5	1/3	7	1/3
Quality	1/3	1	3	1/5	3	1/5
Equity	1/5	1/3	1	1/5	1	1/7
Efficiency	3	5	5	1	3	1
Benefit	1/7	1/3	1	1/3	1	1/7
Cost	3	5	7	1	7	1

Table 2

Relative Importance of Performance Dimensions,

Based upon Multiple Classification Analysis

		Mean Percentages by Dimension					
		Quantity	Quality	Equity	Efficiency	Benefit	Cost
Factor	N						
Grand Mean	640	11%	20%	15%	11%	27%	16%
Method (alpha level)		(.26)	(.32)	(.01)	(.15)	(.63)	(.003)
Social Judgment Theory	88	13	18	16	8	26	18
Simplified Multi. Rating Tech.	331	11	20	12	12	26	19
Analytic Hierar. Process	221	11	20	20	11	29	10
Program (alpha level)		(.87)	(.82)	(.02)	(.45)	(.06)	(.72)
Probation	362	11	20	15	11	26	16
Medicaid	94	12	20	18	11	24	15
Air Pollution	93	11	21	13	10	28	16
Any	91	12	20	15	10	28	14
Subject's Role (alpha level)		(.22)	(.06)	(.003)	(.03)	(.82)	(.004)
Funder	100	10	17	13	12	26	21
Funder surrogate	212	11	22	16	10	27	13
Researcher	72	8	21	19	8	25	18
Researcher surro.	22	12	17	13	13	31	16
Practitioner	155	11	20	15	11	27	16
Practitioner surro.	12	13	18	20	13	27	10
General public	67	13	22	12	11	28	16
Cost Definition (alpha level)		(.40)	(.73)	(.07)	(.72)	(.97)	(.10)
Total cost	368	12	20	18	11	27	13
Cost-effectiveness	272	10	21	12	11	27	18

REFERENCES

Billings, R. S. & Marcus, S. A. Measures of compensatory and noncompensatory models of decision behavior: process tracing versus policy capturing. Organizational Behavior and Human Performance, 1983, 31, 331:352.

Bodily, S. Merging the preferences of interest groups of efficiency and equity of service in the design of police sectors. In R. C. Larson (Ed.), Police deployment: new tools for planners. Lexington, Mass.: Lexington Books, 1978.

Coulter, P. B. Measuring the inequity of urban public services: a methodological discussion with applications. Policy Studies Journal, 1980, 8, 683-697.

Dawes, R. and Corrigan, R. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

Edwards, W. Multiattribute utility for evaluation: structures, uses, and problems. In M. W. Klein and K. S. Teilman (Eds.), Handbook of criminal justice evaluation. Beverly Hills: Sage, 1980.

Edwards, W. Multiattribute utility measurement: evaluating desegregation plans in a highly political context. In R. Perloff (Ed.), Evaluator interventions: pros and cons. Beverly Hills: Sage, 1979.

Edwards, W. Use of multiattribute utility measurement for social decision making. In D. E. Bell, R. L. Keeney, and H. Raiffa, (Eds.), Conflicting objectives in decisions. New York: John Wiley, 1977.

Eils, L. C. & John, R. S. A criterion validation of multiattribute utility analysis and of group communications strategy. Organizational Behavior and Human Performance, 1980, 25, 268-288.

Einhorn, H. J. & McCoach, W. A simple multiattribute utility procedure for evaluation. Behavioral Science, 1977, 22, 270-282.

Hammond, K. R. Externalizing the parameters of quasirational thought. In M. Zeleny (Ed.), Multiple criteria decision making, Kyoto 1975. Berlin: Springer-Verlag, 1976.

Hammond, K. R., Rohrbaugh, J., Mumpower, J, & Adelman, L. Social judgment theory: applications in policy formation. In M. F. Kaplan and S. Schwartz (Eds.), Human judgment and decision processes: applications in problem settings. New York: Academic Press, 1977.

Hershey, J. C., Kunreuther, H. C., & Schoemaker, P. J. H. Sources of bias in assessment procedures for utility functions. Management Science, 1982, 28, 936-954.

Hobbs, B. F. A comparison of weighting methods in power plant siting. Decision Sciences, 1980, 11, 725-737.

Hwang, C. L. & Yoon, K. Multiple attribute decision making methods and appplications: a state-of-the-art survey. Berlin: Springer-Verlag, 1981.

Jones, B. D. Assessing the products of government: what gets distributed? Policy Studies Journal, 1981, 9, 963-971.

Lineberry, R. L. & Welch, R. E. Who gets what: measuring the distribution of urban public services. Social Science Quarterly, 1974,54,700-712.

Ostrom E., Parks, R. B., Percy, S. L. & Whitaker, G. P. Evaluating police organization. Public Productivity Review, 1979, 3, 3-27.

Pitz, G. F., Heerboth, J., & Sachs, N. J. Assessing the utility of multiattribute utility assessments. Organizational Behavior and Human Performance, 1980, 26, 65-80.

Rohrbaugh, J. Improving the quality of group judgment: social judgment analysis and the nominal group technique. Organizational Behavior and Human Performance, 1981, 28, 272-288.

Saaty, T. L. The analytic hierarchy process: planning, priority setting, resource allocation. New York: McGraw-Hill, 1980.

Schoemaker, P. J. H. & Waid, C. C. An experimental comparison of different approaches to determining weights in additive utility models. Management Science, 1982, 28, 182-196.

Stillwell, W. G., Seaver, D. A., & Edwards, W. A comparison of weight approximation techniques in multiattribute utility decision making. Organizational Behavior and Human Performance, 1981, 28, 62-77.

Wallsten T. S. & Budescu, D. V. Encoding subjective probabilities: a psychological and psychometric review. Management Science, 1983, 29, 151-173.

Wilenski, P. Efficiency or equity : competing values in administrtive reform. Policy Studies Journal, 1980-81, 9, 1239-1249.

THE OSPREY COMPANY

INTEGRATING NEW METHODS FOR ANALYZING GROUP DECISION MAKING: SOCIAL JUDGMENT THEORY, FUNCTIONAL FORMS AND RANDOM COEFFICIENT MODELS

by

Gloria A. Grizzle and Ann D. Witte

Working Paper 83-1
March 1983

submitted to Management Science

Prepared under grant 80-IJ-CX-0033 from the National Institute of Justice, U.S. Department of Justice. Views and opinions are those of the author and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

INTEGRATING NEW METHODS FOR ANALYZING GROUP DECISION MAKING:
SOCIAL JUDGMENT THEORY, FUNCTIONAL FORMS
AND RANDOM COEFFICIENT MODELS

Researchers interested in such diverse topics as family and firm decisionmaking and the choice among public policies have developed methods for understanding and improving group decision making. However, the work to date has been fragmented along disciplinary lines and, thus, has not made as much progress as we believe is possible. In this paper, we seek to integrate from several disciplines insights concerning (a) the appropriate functional form for analyzing the multiattribute decision process and (b) the appropriate methods for combining individual preferences. We believe that we provide useful insights on methods of (a) analyzing judgments about the relative merits of alternative public sector programs and (b) identifying the major factors causing disagreement in the public decision making process.

The outline of the paper is as follows. Section one reviews three different approaches to public sector decision making (social welfare functions, multiattribute utility analysis and social judgment analysis). Section two describes our model of public decision making and the way in which we plan to deal with diverse individual valuations of the many attributes of public programs. Section 3 describes the data used to estimate our model. Sections 4 and 5 contain a discussion of the results obtained when estimating the model using individual and group data, respectively. The final section of the paper contains a

summary and conclusions regarding methods of analyzing decision making in the public sector.

METHODS OF ANALYZING PUBLIC DECISIONS

There are presently three major methods of analyzing public decision making: social welfare functions and benefit-cost analyses, multiattribute utility analysis and social judgement analysis. We discuss each of these approaches briefly and at the end of the section indicate how our work relates to each of these techniques.

Economists have long sought to analyze societal decision making and early on developed the concept of a social welfare function. However, Arrow (1951) early showed the impossibility of basing social choice on individual values, the mainstay of economists' analysis of choice, without making explicit individual comparisons. Since Arrow's work, economists have analyzed both normative (e.g., Rawlesian) and actual (e.g., majority voting) methods of social decision making. See Deaton and Muellbauer (1980) and Mueller (1979) for recent surveys. This work has been mainly concerned with theoretical issues and has been little used to study actual decision making. By way of contrast, economists have developed benefit-cost analysis, a rather formidable set of tools, to determine the economically efficient choice among alternative uses of public funds. As is well known, this type of analysis considers mainly efficiency issues and attempts to reduce all, or at least most, benefits and costs to monetary equivalents. See Mishan (1982) for a survey of this literature.

A number of operations researchers and some psychologists interested in practical aids for decision making have developed

multiattribute utility analysis to structure and understand decision making. Researchers using multiattribute utility analysis (MAUT) work with decision makers to determine the attributes of the problem central to the decision at hand and the relative importance of each attribute. For surveys of MAUT, see Keeney and Raiffa (1976); Starr and Zeleny (1977), Edwards (1980) and Hwang and Yoon (1981). This literature has rather thoroughly considered the functional form appropriate for aggregating attributes (linear vs. various non-linear forms) and methods of incorporating uncertainty. In selecting a particular functional form, these researchers frequently use insights from the economics literature on consumer demand.

Analysts using social judgment analysis, mainly social psychologists, have practical interests similar to researchers using multiattribute utility analysis. Whereas researchers using mutiattribute techniques seek to determine the relative importance of each aspect of the decision problem, analysts using social judgment theory seek only overall evaluations of different solutions. Given these overall or holistic judgements, analysts using social judgement analysis seek to determine the implicit valuation of the important attributes of the decision under consideration by regressing the holistic judgement on the known attributes of the different solutions evaluated. See Hammond, et al. (1975) for a description of the technique, and Hammond, et al. (1977), Rohrbaugh (1981), Rohrbaugh and Quinn (1980), Rohrbaugh and Wehr (1978) or Roose and Doherty (1978) for examples of its use.

In this paper, we seek to use insights from all three of the above literatures and to extend them by incorporating fairly recent work from the economics literature dealing with consumer demand and the econometrics literature. Specifically, we utilize a functional form which may be considered a second order approximation to an unknown decision function when analyzing the way in which individuals value the attributes of the decision problem. This approach is quite popular in the economics literature which analyzes consumer demand and production. As far as we are aware this approach has yet to be used to analyze group decision making in a multiattribute setting. We believe it a useful approach because it admits our ignorance of the true form of the decision makers' evaluative function. Further, the form we utilize contains the linear and quadratic forms as special cases and therefore may be used to test assumptions frequently made in the MAUT and social judgment literatures.

Both the MAUT and social judgment literatures have dealt with the problem of developing methods of combining judgments when there are a number of decision makers. MAUT generally combines individual valuations by positing a supra decision maker who acts as a synthesizer or amalgamator of individual preferences or attempts to reconcile differences in evaluation through "shared analysis" of the decision problem. In contrast, social judgement analysis attempts to identify groups of individuals with similar methods of evaluating alternatives. Cluster analysis is the statistical method generally used to identify the groupings. Given these groupings, as an aid to group

reconciliation, analysts display pictorially the way in which the groups value the various attributes of the prob_em.

In our work, we attempted to use the cluster analysis approach of the social judgement analysts. However, we were unable to identify groups of reasonable size having methods of evaluation insignificantly (in a statistical sense) different from one another. Faced with the dilemma of needing to combine information for decision making and with the differences in the way in which atttibutes were valued, we resorted to a fairly recently developed econometric technique, the random coefficient model, which allows the parameters in a regression analysis to vary from individual to individual rather than be a constant and equal parameter for all individuals. This technique permits estimating the mean valuation of attributes for all individuals or groups judging the problem. Further, it allows us to estimate individual valuations and the degree to which valuations differ across individuals. The estimates of mean valuation may prove useful as a basis for compromise while the estimates of the variability of judgment may be useful in identifying the "bones of contention" in the decision problem.

A MODEL OF DECISION MAKING

We are interested in analyzing the way in which individuals and groups evaluate the performance of public agencies and programs. To conduct this type of analysis, it is first necessary to identify the major dimensions that decision makers use in forming their program or agency evaluations. A previous paper (Grizzle, 1981) surveys the performance measurement literature and concludes that the major

dimensions that people use to define performance include: (1) the amount of direct output (e.g., services rendered or regulations enforced) that the agency or program produces; (2) the quality of the program (e.g., the accessibility, reliability, and timeliness of service, and client or public satisfaction); (3) the equity of the program (e.g., the degree to which such factors as income, race and sex affect the distribution of services); (4) the cost of the program (e.g., total costs, cost effectiveness, cost per unit of direct output); and (5) the ultimate impact or social benefits of the program.

Having determined major dimensions or attributes relevant for judging the public program or agency, it is next necessary to decide on the type of judgments to elicit from decision makers. Recall that MAUT would elicit explicit judgments on each attribute while social judgment analysis would elicit holistic judgments of programs having known values of the attributres. We chose the latter approach because it has the ability to capture the complex trade-offs among a fairly large number of attributes.

Once holistic judgments are obtained, it is necessary to determine the valuation of the attributes implicit in those judgments. This valuation requires selecting a mathematical form for the evaluative function. Most practical applications of both MAUT and social judgment analysis have used quite simple functional forms (generally linear or at most quadratic in the attributes) which do not allow the valuation of one attribute to be affected by the amount of another attribute present. The theoretical literature, however, clearly indicates the need to allow for such interaction effects.

Economists have devoted considerable theoretical and empirical work to determining the mathematical form appropriate to analyze individual decision making. They have generally concluded that simple forms such as the linear, log linear, and quadratic imply restrictions on the shape of individual utility functions which do not seem appropriate. See Deaton and Muellbauer (1980) for a discussion. In recent years, economists have increasingly used second order approximations to unknown functional forms in both their analysis of consumer demand and firm production. (For an example from the consumer demand literature, see Christensen, Jorgenson and Lau, 1975.) There are now a number of such second order approximations (see Fuss, McFadden and Mundlak, 1978 for a discussion and comparison), and the analysts' selection among them depends largely on the nature of the problem to be analyzed and the analytic issues which are of primary concern. We chose a generalized quadratic form originally suggested by Lau (1974) because it contains the much-used linear and simple quadratic forms as special cases, thereby allowing us to test explicitly for the appropriateness of these more restrictive forms. The generalized quadratic which we use can be seen as a second order Taylor's series expansion of the holistic evaluation (y) in the attributes (x_is). Specifically the form we utilize to analyze individual valuations is:

$$y = \alpha + \sum_{k=1}^{n} \alpha_k x_k + \sum_{k=1}^{n} \sum_{j=1}^{n} \alpha_{kj} x_k x_j + \varepsilon \quad (1)$$

where the αs are parameters to be estimated, n is the number of attributes and ε is a stochastic error term.

Note that this form contains linear, quadratic and interaction terms in the attributes. One may test for linearity by testing the hypothesis that all α_{jk}s jointly equal zero and for a simple quadratic form by testing the hypothesis that the α_{kj}s equal zero whenever $k \neq j$.

Turning attention to group rather than individual decision making, one faces the problem of selecting a method of combining individual valuations. Most work to date seeks rules which allow explicit and non-stochastic aggregation of these preferences. Examples include the Rawlesian social welfare function of economics, the supra decision maker of MAUT and the cluster analytic techniques of social judgment analysis. We choose an alternative technique which incorporates differences in individual valuations. Specifically, when analyzing group decision making, we utilize a random coefficient model due to Swamy (1970, 1971). This model implies the following equation for individual valuation of a public program or agency:

$$Y_i = \alpha_{oi} + \sum_{k=1}^{n} \alpha_{ki} x_{ki} + \sum_{k=1}^{n} \sum_{j=1}^{n} \alpha_{kji} x_k x_j + \varepsilon_i \qquad (2)$$

where i goes from 1 to D (the number of decision makers). Note that the parameters of this model are allowed to differ for different decision makers. The Swamy model assumes that the attribute valuation for each individual (the αs) can be regarded as a random vector drawn from a probability distribution, with mean $\bar{\alpha}$ and a covariance matrix which we will call Δ. With these assumptions we may write equation (2) as follows:

$$Y_i = (\bar{\alpha}_o + \mu_{oi}) + \sum_{k=1}^{n} (\bar{\alpha}_k + \mu_{ki}) x_{ki} + \sum_{k=1}^{n} \sum_{j=1}^{n} (\bar{\alpha}_{kj} + \mu_{kji}) x_k x_k + \varepsilon_i$$

We are interested in estimating the "average" valuation of attributes (the $\bar{\alpha}$s), the degree to which such valuations vary among decision makers ($E[\mu_i \mu_i']$) and a valuation for each individual (α_is). We can obtain estimates of the mean valuation of attributes and the variation among decision makers in these valuations using a Generalized Least Squares (GLS) technique. (See Judge, et al. (1980) for a description.) Lee and Griffiths (1979) have developed an unbiased estimator for the individual coefficients which can be viewed as an estimator of the mean response plus a predictor of the individual variance from this mean ($\hat{\mu}_i$).

THE DATA

We estimate our model using two distinct data sets. The first data set, for which we give results here, consists of data on the holistic evaluations of eighty separate agency profiles by graduate students in a public budgeting course. Many students were currently working in government. Prior to asking for the holistic evaluations, students were given lectures on performance measurement to familiarize them with the basic nature of the task to be completed. In two separate sessions the students were given instructions for completing the exercise, descriptions of program attributes, (see Table 1) and the hypothetical performance of different agencies, scored in terms of these attributes. The scores for an attribute ranged from 1 (indicating poorest performance) to 10 (best performance). Table 2 displays the profiles of eight of these agencies. During the first session, the students rated the performance of 40 agencies on a scale of 0 (poorest performance) to twenty (best performance). Two weeks

later they rated the performance of forty additional agencies, using the same scale.[1] The second data set is similar to the first and contains evaluation by 12 graduate students in a program evaluation course. We used this second data set to corroborate our results. Results using this second data set are similar to those using the first and are not discussed here. However, the similarity of results using this second data set gives additional support for our conclusions.

THE INDIVIDUAL RESULTS

Space and the reader's patience do not allow presenting individual results for all 33 individuals in our first data set. To give the flavor of our findings we present in the first half of Table 3 results obtained using ordinary least squares (OLS)[2] for five randomly selected respondents. We briefly summarize below results for all 33 respondents.

As a whole our results appear quite reasonable. Our model explains a statistically significant[3] amount of the variation in individual valuation of the hypothetical agencies. The coefficient of determination (R^2) for the individual models ranged from .53 to .95.

Turning from general measures of the "goodness of fit" of the models to results for individual variables, one notes diversity. These results strongly suggest that individual methods of valuing performance vary widely. Diversity appears to be more often reflected in differing perceptions of which attributes are important than in widely differing valuations of particular attributes.

Wishing to determine if it was possible to simplify our model, we examined the t-statistics for the coefficients on all variables for

all respondents. We found that all of the variables in the model had large values of the t-statistic for at least some respondents, suggesting that they should all be retained in a model seeking to explain overall valuations of agency performance.[4] Of the attributes considered, more individuals appear to value independently[5] the ultimate impact (BENEF) and quantity of program output. A large number of respondents also used unit cost in their assessment of agency performance although the coefficient on this variable was usually of rather low significance. The equity and quality of agency performance significantly affect the evaluation of relatively few of the 33 respondents, but feelings concerning the importance of these aspects of agency performance were often quite strong. When both linear and quadratic terms affected individual evaluation, the coefficients on those variables indicated that the marginal value of an attribute decreases as the amount of the attribute increases. This result is encouraging because it conforms to the theoretical expectation of diminishing marginal utility.

Turning to the interaction terms, one is struck by their importance in determining individual valuations of agency performance. The coefficients on all interaction terms were significantly different from zero for at least three respondents.[6] The coefficients on the interaction term for quantity of output and total costs (QUANT*TOTSCT) and quality of output and total (QUAL*TOTSCT) were significant for more respondents than any other interaction term. When significant, the coefficient on the interaction term between quality of output and total costs was always positive, indicating that almost half the

respondents value higher quality of agency output and lower agency total costs when performance on these two attributes improve jointly rather than singly. For 11 of the 14 respondents who significantly valued the quantity of agency output and total costs jointly, the coefficient on this interaction term was negative. This negative sign indicates that increases in agency output and increases in total costs were most valued when they occurred together. Table 4 summarizes results for all interaction terms.

Possessed of a large number of diverse results, we next tried a number of methods to identify groups with similar ways of valuing agency performance. We began by testing to see if we could accept the null hypothesis that the coefficients on all variables for all individuals were insignificantly different using an F-statistic. Results indicated that there were major differences in methods of valuing public programs and that it was not possible to pool all members of the sample to obtain a "consensus" valuation.[7]

We next tried to identify subgroups for which methods of valuation were insignificantly different using two different approaches. First, for each individual we identified a reduced specification containing only variables which appeared to be important in performance evaluation for that individual. We used a modified version of Thiel's residual variance criterion (Thiel, 1960, pp. 210-215) to identify "relevant" variables. Specifically, we selected the model which minimized the estimated standard error of the disturbance subject to the condition that the coefficients on all deleted variables be jointly as well as individually insignificantly different

from zero. The seco d half of Table 3 contains the reduced specifications which resulted for the five respondents whom we selected randomly. We next identified subgroups that contained similar variables in their reduced specification and conducted F-tests to determine whether or not we were justified in pooling information for individuals in these subgroups to obtain subgroup methods of valuation. Results indicated that we were not justified in pooling any of the subgroups identified.

We next turned to cluster analytic procedures to identify groups with similar methods of evaluation.[8] We developed clusters based on (1) the correlations of the individual ratings of agency performance; (2) the coefficients of the fully specified model; and (3) the t-ratios of the fully specified model. In no instance were we able to identify reasonably sized subgroups for which we could accept the null hypothesis that subgroup members had similar methods of valuing agency performance.

THE GROUP RESULTS

Having failed to identify subgroups of individuals with similar methods of valuing agency performance, we decided to use data for all individuals to estimate a random coefficient model. Recall that this model allows the coefficients of variables to vary across individuals. Table 5 reports the estimates of the mean coefficient, the "t-ratio," which tests for the significance of the mean coefficient,[9] the estimated standard deviation of the coefficients across individuals, and the ratio of the estimated standard deviation ($\hat{\sigma}_b$) to the mean coefficient ($\hat{\bar{b}}$).

Consider first only the linear term. If we examined only these terms we would conclude that the group judged programs primarily based on the quantity of output they produce (QUANT), the total cost of producing the output (TOTCST), the unit costs of producing the output (UNITCST) and the ultimate impact of the program (BENEF). We would conclude that neither differences in the quality of agency output (QUAL) nor in the equity with which output was distributed (EQUITY) significantly affected valuation.[10]

However, when we consider quadratic and interaction terms, we find that the quality of output and the equity of its distribution significantly affect valuation. Specifically, increases in output quality are positively valued as long as they are associated with equitable distribution and moderate or low total agency cost. If output is distributed inequitably and/or costs are high, increases in quality are likely to lead to lower levels of evaluation for program output. Increases in the equity with which output is distributed will be positively valued as long as both the quantity and quality of output are at reasonable levels.

Returning to the average way in which the groups' valuation of agency performance changed with increased program output, unit cost, benefit and total cost, we find that valuation of all of these items depends on more than simply the magnitude of the item under consideration. The simplest relationship between a variable and agency performance evaluation occurs for ultimate program impact (BENEF). Our results indicate that agency performance valuation goes up with increased program impact but at a decreasing rate (i.e., the

coefficient on $BENEF^2$ is negative and significantly different from zero). Similarly, the valuation of unit costs decreases as the magnitude of these costs go down. However, the valuation of unit cost also depends on the level of agency output. As the quantity of program output goes up the valuation of decreases in unit costs declines. How decreases in the total cost of agency operation are valued depends upon both the quantity and quality of program output. Higher valuations are associated with lower levels of output, but higher quality of output. Finally, the way in which increases in the quantity of program output are valued is quite complex. Increased quantity of output is valued at a decreasing rate as the quantity of output goes up. However, the valuation of increased output also depends on the equity and costs of output. Higher levels of output are more highly valued if the output is equitably distributed at relatively low unit and total cost.

Overall, our results indicate that members of the group independently value only increases in ultimate program impact. How they value the other five performance dimensions depends upon not only how well an agency performs in terms of single dimensions but also in terms of how these dimensions associate with each other. For example, lower costs may not be posiively valued when the quantity of output is very high and the quality very low.

Turning from the mean coefficients of the variables to their variability among group members, we find that individuals in the group differed most in the way they valued output quality, equity and total costs.[11] Thus, our results lead us to believe that group disagreement on agency performance will stem mainly from differences in valuation

of the quality of agency output, equity with which output is distributed and total agency spending. These results do not seem unreasonable. Least difference in valuations surround program impact and the unit cost of agency output.

Finally, using a test statistic suggested by Swamy (1970), we test to see if our assumption of a random vector of coefficients is valid. We strongly reject the hypothesis that there are no differences in coefficients among individuals[12] and, thus, conclude that the random coefficient model is an appropriate model for representing the way in which group members value agency performance.

SUMMARY AND CONCLUSIONS

In this paper, we have considered the problem of modeling individual and group judgments. We suggest that current methods often use functional forms which are too simple to reflect methods of individual decision making in important and complex situations and suggest using a second order approximation to capture the complexity of judgment in such situations. Turning next to the problem of aggregating individual preferences in order to obtain group valuations, we suggest use of a method, the random coefficient model, that specifically recognizes the heterogeneity of individual valuations. To make our suggestions more concrete and to explore their usefulness, we next use these methods to explore the way in which a group of individuals evaluate agency performance. We find that the traditionally used linear and quadratic forms are too simple to mirror accurately the judgment process for our group. Interaction terms are

important in determining group members' evaluations of agency performance.

Next, we estimate a random coefficient model of valuation for the group as a whole. We test for the appropriateness of this model and find that a random rather than a fixed coefficient approach more accurately reflects the nature of group valuation, i.e., individual methods of valuation are too diverse to be adequately represented by a single set of parameters. We find that group valuation of agency performance is strongly and independently affected by the ultimate impact the agency's activities have on clients and society as a whole. Valuation is also independently, but less strongly, affected by the quantity of services the agency produces and the unit and total costs for which these services are produced. The quality of agency services and the equity with which services are distributed only positively affect agency valuation when found in combination with other agency attributes (i.e., only coefficients on interaction terms in these variables are positive and significant). For example, our results indicate that an increase in the equity with which an agency's output is distributed will only positively affect the group's valuation of agency performance if this increase in equity occurs in an agency with reasonable levels of services in terms of both quantity and quality. The estimated standard deviations on the coefficients of our model allow us to identify factors for which there are extensive intergroup differences in valuation. We find the greatest group differences in valuation for the quality of output, the equity of its distribution and total agency spending.

We conclude that relatively complex functional forms are required to reflect individual and group decision making proceses adequately. Specifically, it appears that the valuation of one agency or program attribute is affected by the level of other attributes of the program or agency. Further, in spite of rather extensive efforts, we were unable to identify subgroups of reasonable size that had similar (in a statistical sense) methods of valuing agency performance. Thus, we conclude that it is appropriate to use methods which specifically recognize the heterogeneity of individual preferences when evaluating public agencies or programs. We illustrate the use of one such technique, the random coefficient model.

FOOTNOTES

[1]We tested to see if the methods of valuation for individuals varied in the two sessions and were able to accept the null hypothesis that they did not vary. Thus, we pool valuation of all 80 agency profiles.

[2]We noted that our dependent variable was truncated at 0 and 20 and thus, carefully examined the plot of residuals from the individual OLS regressions to determine if this truncation had resulted in violation of OLS assumptions. As there was no significant pile-up of observations at either 0 or 20, these plots did not reveal the violations normally associated with truncated dependent variables (non-normality and non-zero mean for the residuals). Further, these plots indicated no consistent violation of other OLS assumptions although for a few individuals the variance of the residuals for middle range residuals were somewhat higher than for either high or low values for the residuals (i.e., there appeared to be a moderate degree of heteroskedasticity for some individuals). As this problem was neither marked nor pervasive, we chose to ignore it.

[3]Specifically, the F-statistic which tests for the significance of the total model's explanatory power was greater than the .01 critical point in all instances.

[4]Specifically, the absolute value of the t-statistic testing the significance of the coefficient on individual variables was greater than 1.28 for at least one individual for all variables. Most model selection criteria (see Judge, et al., 1980, for a discussion) suggest using low values for the t-statistic when deciding whether or not to retain a variable in a model.

[5]When we use the term "independently value," we are considering only the coefficients on linear and quadratic terms.

[6]In this section, which considers model specification, we will judge the coefficient on a variable to be significant if the absolute value of its t-ratio is greater than 1.28.

[7]For a discussion of appropriate tests, see Maddala (1977). The value of the F-statistic, which is distributed $F_{896, 1716}$ under the null hypothesis, was 4.18 which indicates that we cannot accept the null hypothesis (methods of evaluation are similar) at normal levels of statistical significance (e.g. = .01 or .05).

[8]See Hudson and Associates (1982) for a discussion of various methods for identifying similar individuals. We used a method which sought to minimize the difference among subgroup members. The actual procedure utilized is contained in the SAS package of computer programs.

[9]This statistic is distributed N(0,1) under the null hypothesis that the mean coefficient is insignificantly different from zero. It should be noted that we obtain our estimate of the standard deviation of the coefficients by assuming tht the relevant covariance matrix is diagonal and by adjusting the original covariance matrix which was not nonnegative definite in a manner suggested by Judge et al. (1981, p. 350).

[10]Note, in this section, we judge variables to be significantly related to performance valuation if the coefficient on the variable would be judged to be significantly different from zero at the ten percent level, two tailed test.

[11]We consider only the variables with significant mean coefficients and rely on the value of $\hat{\sigma}_b/\hat{\bar{b}}$ to determine the degree of variability.

[12]The value of the test statistic, which is distributed $\chi^2_{.996}$ under the null hypothesis of fixed coefficients, is 4132.93, which clearly indicates that we cannot accept the null hypothesis at normal levels of statistical significance.

BIBLIOGRAPHY

ARROW, KENNETH. (1951) Social Choice and Individual Values. New York: John Wiley.

BRENNAN, TIM. (1980) Multivariate Taxonomic Classiciation for Criminal Justice Research. Washington, D.C.: Final Report on Project No. 78-NJ-AX-0065 of the National Institute of Justice.

CHRISTENSEN, L. R., D. W. JORGENSON and L. J. LAU. (1975) "Transcendental logarithmic utility functions." American Economic Review 65:367-383.

DEATON, ANGUS and JOHN MUELLBAUER. (1980) Economics and Consumer Behavior. Cambridge, England: Cambridge University Press.

EDWARDS, WARD S. (1980) "Multi-attribute utility for evaluation," pp. 177-216 in Malcolm W. Klein and Katherine S. Teilmann (eds.) Handbook of Criminal Justice Evaluation. Beverly Hills, CA.: Sage.

FUSS, MELVYN, DANIEL MCFADDEN and YAIR MUNDLAK. (1978) "A survey of functional forms in the economic analysis of production," pp. 219-268 in Melvyn Fuss and Daniel McFadden (eds.) Production Economics: A Dual Approach to Theory and Applications, Vol. 1. Amsterdam: North-Holland.

HAMMOND, KENNETH R., et al. (1975) "Social judgment theory," pp. 271-312 in Martin F. Kaplan and Steven Schwartz (eds.), Human Judgment and Decision Processes. New York: Academic Press.

HAMMOND, KENNETH R., et al. (1977) "Social judgement theory: applications in policy formation," pp. 1-29 in Martin F. Kaplan and Steven Schwartz (eds.), Human Judgment and Decision Processes in Applied Settings. New York: Academic Press.

HUDSON, HERSCHEL C. and Associates. (1982) Classifying Social Data. San Francisco: Jossey-Bass Publishers.

HWANG, CHING-LAI and YOON, KWANGSUN. (1981) Multiple Attribute Decision Making Methods and Applications: A State-of-the-Art Survey. Berlin: Springer-Verbeg.

GRIZZLE, G. A. (1981) "A manager's guide to the meaning and uses of performance measurement." American Review of Public Administration 15:16-28.

KENDALL, MAURICE and STUART, ALAN. (1976) The Advanced Theory of Statistics. Volume 3. London: Charles Griffin.

KEENEY, RALPH L. and RAIFFA, HOWARD. (1976) Decisions with Multiple Objectives: Preferences and Trade-offs. New York: John Wiley.

JOHNSON, STEPHEN C. (1967) "Hierarchical clustering schemes." Psychometrika xxxii:24-254.

JUDGE, GEORGE C., et al. (1980) The Theory and Practice of Econometrics. New York: John Wiley.

MADDALA, G. S. (1977) Econometrics. New York: McGraw-Hill.

MISHAN, E. J. (1982) Cost-Benefit Analysis: New and Expanded Edition. London: George Allen & Unwin.

MUELLER, DENNIS C. (1979) Public Choice. Cambridge, England: Cambridge University Press.

LEE, L. F. and GRIFFITHS, W. E. (1979) "The prior likelihood and best linear unbiased prediction in stochastic coefficient linear models," Working Paper in Econometrics and Applied Statistics, No. 1. Armidale, Australia: University of New England.

LAU, LAWRENCE J. (1974) "Comments on applications of duality theory," pp. 176-199 in M. D. Intriligator and D. A. Kandrick (eds.), Frontiers in Quantitative Economics, Vol. II. Amsterdam: North-Holland.

ROOSE, JACK E. AND MICHAEL E. DOHERTY. (1978) "A social judgment theoretic approach to sex discrimination in faculty salaries." Organizational Behavior and Human Performance 22:193-215.

ROHRBAUGH, JOHN (1981) "Improving the quality of group judgment: social judgement analysis and the nominal group technique." Organizational Behavior and Human Performance 28:272-288.

ROHRBAUGH, JOHN and QUINN, ROBERT (1980), "Evaluating the performance of public organizations: a method for developing a single index." Journal of Health and Human Resources Administration 2,3:343-354.

ROHRBAUGH, JOHN and WEHR, PAUL. (1978) "Judgment analysis in policy formation: a new method for improving public participation." Public Opinion Quarterly 42:521-532.

STARR, MARTIN K. and ZELENY, MILAN, eds. (1977) "Multiple criteria decision making." TIMS Studies in the Management Sciences 6:5-30 and 59-90.

SWAMY, P. A. V. B. (1971) Statistical Inference in Random Coefficient Regression Models. New York: Springer-Verlag.

SWAMY, P. A. V. B. (1970) "Efficient inference in a random coefficient regression model." Econometrica 38:311-323.

THIEL, H. (1960) Economic Forecasts and Policy. Amsterdam: North-Holland.

Table 1

Definitions of Performance Dimensions and Acronyms

Quantity of output (denoted QUANT) refers to the amount of a program's direct product, i.e. the services rendered or regulations enforced.

Examples: Number of children screened
Number of noncomplaince citations delivered
Number of prisoners placed on parole
Number of miles of street paved

Quality of program (denoted QUAL) refers to how well the program is working and encompasses a number of attributes, including accessibility of the service to the client, the degree to which outputs are reliable and valid, the client's and public's satisfaction with the service received or the regulations enforced, timeliness of the service, and cost to the client (both economic and psychological) of receiving the service.

Examples: Average length of time between referral and diagnosis
Percentage of health problems not found during screening
Percentage of complaints about pollution-control violations followed up within one week
Average waiting time for clients
Percentage of clients located within an hour's ride of the service center

Equitable distribution of outputs (denoted EQUITY) refers to how services or the enforcement of regulations are distributed among people. Common ways of breaking down service delivery in order to look at the equity of distribution include geographic area, sex, race, age, education, economic status, and extent of need.

Examples: Percentage of applicants served, by each county in the state
Percentage of job placements, by age group
Percentage of street paved, by census tract

Cost per unit of output (denoted UNITCST) is obtained by dividing program total cost by quantity of output. Note a value of 10 indicates a low level of unit costs and a value of 1 a high level.

Examples: Cost per child screened
Cost per noncompliance citation delivered
Cost per prisoner placed on parole
Cost per mile of street paved

Benefit to society (denoted (BENEF) refers to the effect or impact of the program upon clients who were directly served or other groups who were indirectly affected as a result of the program's outputs.

Examples: The dollar value of damage to agriculture avoided because of improved air quality
The number of children with vision problems that have been corrected because of the program
Reduction in crimes committed due to supervision of probationers and parolees
Increase in the probability that former clients will be healthy through subsequent phases of the life cycle.

Total program cost (denoted TOTCST) refers to the amount of money spent on the program for all purposes, such as salaries for program staff, supplies, travel, and equipment. Note that a value of 10 indicates high performance in terms of total cost and a value of 1 indicates low performance.

Table 2

Examples of Performance Profiles

1

Quantity of Output	XXXXXXXXXX
Quality of Program	XX
Equitable Dist.-Outputs	XXXXXXXXX
Unit Cost - Output	XXXX
Benefit to Society	XXXXXX
Total Program Cost	XXXXXXXXX

2

Quantity of Output	XXXXXXXX
Quality of Program	XXXXX
Equitable Dist.-Outputs	X
Unit Cost - Output	XXXXXXXXX
Benefit to Society	XXXXXXX
Total Program Cost	XXXXXX

3

Quantity of Output	XX
Quality of Program	XXXXXX
Equitable Dist.-Outputs	XXXXXXX
Unit Cost - Output	XXX
Benefit to Society	XXXX
Total Program Cost	XXXXXX

4

Quantity of Output	XXXX
Quality of Program	XXXXXXX
Equitable Dist.-Outputs	XXXXXXXX
Unit Cost - Output	XX
Benefit to Society	XXXXXXXXXX
Total Program Cost	XXX

5

Quantity of Output	XXXXXXX
Quality of Program	XXXXXX
Equitable Dist.-Outputs	XXXXXX
Unit Cost - Output	XXXXXXXXXX
Benefit to Society	XXX
Total Program Cost	XXXXXXXXX

6

Quantity of Output	XXXX
Quality of Program	XXXXXXX
Equitable Dist.-Outputs	XXXXX
Unit Cost - Output	XX
Benefit to Society	XXXXXX
Total Program Cost	XX

7

Quantity of Output	XXXXXXX
Quality of Program	XXXXXXXXXX
Equitable Dist.-Outputs	X
Unit Cost - Output	XXXXXXXXXX
Benefit to Society	XXXXXXXXXX
Total Program Cost	XXXXXXXX

8

Quantity of Output	XXXXXXXX
Quality of Program	XXXXXX
Equitable Dist.-Outputs	XXXXXXXX
Unit Cost - Output	X
Benefit to Society	XXXXXXXXX
Total Program Cost	XXXXX

Table 3

Results for Selected Individuals Using Individual Judgments

(t-ratios in Parenthesis)

Independent Variable → / Respondent ↓	Intercept	Linear terms: QUANT	QUAL	EQUITY	UNITCST	BENEF	TOTCST	Quadratic terms: $(QUANT)^2$	$(QUAL)^2$
	(Selected Results For the Fully Specified Model)								
6	-5.545 (1.23)	-0.437 (-0.63)	0.156 (0.22)	1.454** (2.14)	1.092 (1.61)	0.979 (1.32)	-0.304 (-0.50)	0.051 (1.05)	-0.112** (-2.32)
10	-2.244 (-0.30)	0.859 (0.77)	0.704 (0.62)	0.846 (0.77)	0.942 (0.87)	0.879 (0.72)	-0.390 (-0.38)	-0.108 (-1.40)	-0.083 (-1.06)
17	-6.643 (-1.52)	1.226* (1.82)	0.934 (1.36)	1.037 (1.58)	1.112* (1.70)	1.416* (1.97)	0.042 (0.072)	-0.022 (-0.48)	-0.107** (-2.29)
24	-0.504 (0.12)	0.873 (1.30)	-0.106 (-0.16)	-0.197 (-0.30)	0.938 (1.44)	0.972 (1.36)	0.546 (0.93)	-0.135*** (-2.93)	-0.067 (-1.44)
28	9.791* (1.98)	-0.583 (-0.76)	-0.329 (-0.42)	-0.202 (-0.27)	0.422 (0.57)	-0.275 (-0.34)	-0.184 (-0.27)	0.064 (1.21)	-0.059 (-1.12)
	Selected Results for the Reduced Specification								
6	-6.7180** (-2.51)			0.8451 (1.64)	1.1755** (2.33)	1.2440** (2.35)			-0.0728** (-2.77)
10					1.3413*** (5.48)			-0.0412 (-1.13)	
17	-5.3310** (-2.29)	0.2901* (1.87)	0.6997* (1.70)	1.0670*** (2.71)	1.2594*** (3.14)	1.785*** (3.74)			-0.0906** (-2.36)
24		1.0521** (2.57)			0.9445** (2.43)	0.5641** (2.10)		-0.1163*** (-3.54)	-0.0598*** (-2.85)
28	7.3474*** (10.09)								-0.0498** (-2.25)

Table 3 (cont'd)

Independent Variable / Respondent	Quadratic terms (EQUITY)^2	(UNITCST)^2	(BENEF)^2	(TOTALCST)^2	QUAN*QUAL	QUANT*EQUITY	QUAN*UNITCST
	Selected Results For the Fully Specified Model						
6	-0.093**	-0.085	-0.064	0.004	-0.021	0.053	-0.015
	(-2.02)	(-1.66)	(-1.33)	(0.09)	(-0.49)	(1.35)	(-0.33)
10	-0.044	-0.025	-0.126	-0.015	0.014	0.005	-0.053
	(-0.60)	(-0.30)	(-1.62)	(-0.21)	(0.21)	(0.09)	(-0.72)
17	-0.051	-0.092*	-0.064	0.012	-0.047	0.035	-0.038
	(-1.15)	(-1.86)	(-1.37)	(0.28)	(-1.15)	(0.93)	(-0.87)
24	-0.056	-0.056	-0.042	-0.042	0.016	0.152***	0.069
	(-1.27)	(-1.36)	(-0.91)	(-1.00)	(0.38)	(4.07)	(1.57)
28	-0.064	-0.042	0.056	0.023	0.020	0.055	-0.014
	(-1.27)	(-0.75)	(1.05)	(0.48)	(0.42)	(1.28)	(-0.28)
	Selected Results for the Reduced Specification						
6	-0.0521	-0.690*	-0.0633			0.0314	
	(-1.43)	(-1.81)	(-1.61)			(1.26)	
10			0.0551***				0.2241***
			(3.92)				(4.09)
17	-0.0414	-0.0767**	-0.0905**				-0.0356
	(-1.23)	(-2.17)	(-2.58)				(-1.47)
24	-0.0653***					0.1477***	0.0394
	(3.29)					(4.71)	(1.27)
28	-0.0550***		0.0377*				
	(-3.47)		(1.92)				

Table 3 (cont'd)
Interaction terms

Independent Variable Respondent	QUANT*BENEF	QUANT* TOTALCST	QUAL* EQUITY	QUAL* UNITCST	QUAL* BENEF	QUAL* TOTALCST	EQUITY* UNITCST	EQUITY* BENEF	EQUITY* TOTALCST
Selected Results for the Fully Specified Model									
6	0.057 (1.12)	-0.025 (-0.57)	0.074 (1.53)	0.065 (1.47)	0.025 (0.58)	0.097** (2.44)	-0.073* (-1.93)	-0.063 (-1.53)	-0.018 (-0.40)
10	0.028 (0.35)	0.155** (2.20)	-0.008 (-0.10)	-0.048 (-0.68)	0.073 (1.07)	0.058 (0.91)	0.022 (0.36)	-0.038 (-0.58)	0.029 (0.40)
17	-0.037 (-0.76)	-0.089** (-2.10)	0.004 (0.09)	0.040 (0.93)	0.008 (0.20)	0.070* (1.83)	-0.000 (-0.01)	-0.053 (-1.33)	-0.014 (-0.33)
24	0.015 (0.31)	-0.055 (-1.30)	0.053 (1.14)	0.007 (0.17)	0.157*** (3.82)	-0.003 (-0.07)	-0.079** (-2.17)	0.047 (1.18)	0.023 (0.54)
28	-0.021 (-0.38)	-0.076 (-1.58)	0.137** (2.57)	0.017 (0.34)	0.083* (1.78)	0.006 (0.14)	-0.020 (-0.48)	-0.009 (-0.21)	0.007 (0.14)
Selected Results for the Reduced Specification									
6	0.0457 (1.66)		0.0745*** (2.67)	0.0616* (1.89)		0.0560*** (3.10)	-0.0620* (-1.99)	-0.0466 (-1.43)	
10									
17								-0.0471 (-1.51)	
24		-0.0618** (-2.64)	0.0560** (2.11)		0.1404*** (4.45)		0.0742*** (-2.83)	0.0424 (1.41)	
28	0.0610* (1.80)	-0.0346* (1.93)	0.1126*** (4.16)						

Table 3 (cont'd)

Independent variable Respondent	UNITCST* BENEF	UNITCST* TOTALCST	BENEF* TOTALCST	R^2	F	N
		Selected Results for the Fully Specified Model				
6	0.035 (0.74)	0.046 (1.24)	-0.001 (-0.02)	0.68	4.05***	80
10	0.096 (1.27)	-0.057 (-0.96)	-0.016 (-0.27)	0.54	2.18***	79
17	0.021 (0.45)	0.034 (0.94)	0.035 (0.95)	0.67	3.84***	80
24	-0.069 (-1.52)	-0.024 (-0.67)	-0.065* (-1.80)	0.85	11.03***	80
28	0.021 (0.40)	0.042 (1.04)	0.044 (1.07)	0.70	4.43***	80
		Selected Results for the Reduced Specification				
6				0.64	8.41***	80
10		-0.1381*** (-2.79)		0.86	92.53***	79
17				0.60	8.37***	80
24	-0.0598** (-2.51)		0.0400 (-1.64)	0.96	100.92***	80
28		0.0327** (2.78)	0.0392* (1.78)	0.66	16.88***	80

* Indicates that a coefficient is significantly different from zero at the .10 level, two-tailed test
** " " " " " " " " " " " " .05 " " "
*** " " " " " " " " " " " " .01 " " "

Table 4

Results for Interaction Terms Using Individual Judgments

Variable	Number of Respondents for Whom Coefficient on Variable was Significant (i.e. |t - ratio| > 1.28)	Number of Significant Coefficients With Positive Sign	Number of Significant Coefficients With Negative Sign
QUANT*QUAL	7	6	1
QUANT*UNITCST	5	2	3
QUANT*BENEF	3	1	2
QUANT*TOTCST	14	3	11
QUANT*EQUITY	8	7	1
QUAL*UNITCST	4	2	2
QUAL*EQUITY	6	6	0
QUAL*BENEF	7	6	1
QUAL*TOTCST	12	12	0
EQUITY*UNITCST	4	2	2
EQUITY*BENEF	8	5	3
EQUITY*TOTCST	6	3	3
UNITCST*BENEF	7	0	7
UNITCST*TOTCST	6	5	1
BENEF*TOTCST	6	4	2

Table 5

Results Obtained Using the Random Coefficient Model

Variable	Coefficient $(\hat{b})$	"t-ratio"	Standard Deviation of Coefficient $(\hat{\sigma}_b)$	$\frac{\hat{\sigma}_b}{\hat{b}}$
Intercept	-3.0787	-1.90	7.88	-2.56
QUANT	0.7434	3.43	1.00	1.34
QUAL	0.2445	1.39	0.67	2.74
EQUITY	0.1733	1.02	0.65	3.75
UNITCST	0.6557	3.26	0.90	1.37
BENEF	1.2138	5.09	1.12	0.92
TOTCST	0.3060	1.74	0.76	2.48
$(QUANT)^2$	-0.0277	-1.97	0.08	-2.89
$(QUAL)^2$	-0.0250	-2.17	0.06	-2.40
$(EQUITY)^2$	-0.0295	-2.07	0.04	-1.36
$(UNITCST)^2$	-0.0380	-2.21	0.06	-1.58
$(BENEF)^2$	-0.0535	-3.48	0.07	-1.31
$(TOTALCST)^2$	-0.0129	-1.05	0.05	-3.88
QUANT*QUAL	0.0111	0.96	0.05	4.50
QUANT*EQUITY	0.0236	2.15	0.05	2.11
QUANT*UNITCST	-0.0217	-1.89	0.04	-1.84
QUANT*BENEF	0.0138	1.09	0.05	3.62
QUANT*TOTALCST	-0.0276	-1.80	0.07	2.54
QUAL*EQUITY	0.0306	2.35	0.05	1.63
QUAL*UNITCST	0.0020	0.19	0.04	20.00
QUAL*BENEF	0.0227	1.44	0.08	3.52
QUAL*TOTALCST	0.0216	1.90	0.05	2.31
EQUITY*UNITCST	0.0006	0.06	0.05	83.33
EQUITY*BENEF	0.0049	0.40	0.06	12.24
EQUITY*TOTALCST	0.0021	0.17	0.05	23.81
UNIT*BENEF	-0.0096	-0.77	0.05	5.21
UNIT*TOTALCST	0.0088	0.84	0.04	4.55
BENEF*TOTALCST	-0.0046	-0.38	0.06	13.04

END