From: Computers, Environment and Urban Systems, Vol. 5, 1980

Comput. Environ. Urban Systems, Vol. 5, pp. 137 to 154, 1980
Printed in Great Britain

0198-9715/80/050137-18$02.00/0
Pergamon Press Ltd

Reprinted in: Ernest R. House et al., Evaluation Studies Review Annual, Vol. 7, pp. 585-602. Beverly Hills: Sage, 1982.

USING CLIENTS TO EVALUATE PROGRAMS

BRIAN STIPAK

Institute of Public Administration, 211 Burrowes Building, The Pennsylvania State University, University Park, PA 16802, U.S.A.

Abstract—Client surveys can provide valuable information for monitoring and evaluating public programs. However, the widespread use of measures of clients' satisfaction and of clients' subjective evaluations, without an appreciation of the complications of interpretation and analysis, will set back rather than advance the methodology of program evaluation. This paper examines the important issues concerning the use of client-derived information in program monitoring and evaluation, and critically reviews existing research and current practices. Finally, the paper offers a number of recommendations, including six general rules for using client evaluation and satisfaction ratings.

USING CLIENTS TO EVALUATE PROGRAMS

PROGRAM evaluation is increasingly turning to the clients of public programs for measures of program performance. A growing concern with public sector efficiency has helped foster a consumer perspective, in which citizens are seen not merely as passive recipients of agency services provided according to professional standards, but rather as discriminating consumers who make the final evaluation of program effectiveness. Whereas official records were once viewed as the only source of 'objective' information, program professionals are now recognized as having their own self-interest and biases, which are reflected in agency records. According to this new perspective, clients can therefore provide more valid and less biased information than official records for evaluating public programs.†

Complementing the growth of a consumer perspective in program evaluation, social indicator research (e.g. [1, 2]) has developed measures of people's satisfaction and assessments of their lives to augment traditional economic indicators. This approach arises from a desire to change from an overriding concern with economic prosperity to a greater concern with sense of well-being. Since this approach considers quality of life as ultimately a subjective experience, it attempts to directly measure the subjective dimensions of life quality, rather than relying on typically available objective measures. Analogously, this approach suggests that program evaluators should measure the satisfaction and assessments of program clients, and not just rely on so-called objective data generated by program professionals.

These arguments for using clients to evaluate public programs require close scrutiny. As this paper will show, client evaluations can contribute to program evaluation only under some conditions, and even then usually require careful statistical analysis. The widespread use of client evaluations in evaluation research, without an appreciation of the complications of interpretation and analysis, will set back rather than advance the methodology of program evaluation. This paper will examine the most important issues concerning the use of client-derived performance measures in evaluation research. The paper's scope does not include measuring citizens' preferences for use in policy making, but rather is limited to using client-derived information for measuring program performance or effectiveness. However, the scope does encompass not only clients of human service and social welfare programs, but also clients receiving benefits from services provided by any public agency. For example, citizens living in a municipality are clients, as the term is used in this paper, of all city departments that provide municipal services to the general public.

† See Bush and Gordon (1978, pp. 767–776 [3]) for arguments why client-derived information may be more valid and less biased than information derived from agency records.

TYPES OF CLIENT-DERIVED PERFORMANCE MEASURES

Several types of client-derived information can be used for measuring program performance or effectiveness. Performance measures can be objective or subjective. Some subjective measures are intrinsically important measures of outcomes and others are not. Among subjective measures that are not intrinsically important, a useful distinction is between general vs specific client evaluations. The following sections will examine each of these distinctions.

Objective vs subjective measures

Considerable confusion surrounds the distinction between objective and subjective measures. Nunnally (1975, pp. 107–109 [4]) identifies four different reasons client evaluations of programs have been called subjective:

(1) lack of observable evidence to verify the client's stated evaluation
(2) difficulty in instructing clients how to give an evaluation on a rating scale
(3) instability of client evaluations over time and across clients
(4) artificial influences on how clients make evaluative ratings.

Reason (3) concerns reliability, and reasons (2) and (4) concern possible threats to the validity of measurement. Since problems of reliability and validity plague all performance measures, reasons (2–4) do not clearly differentiate subjective from objective measures. Reason (1), however, concerns an important and useful distinction for evaluation research. A client's evaluation of a program on a rating scale, or a client's expressed satisfaction with a service, measures a psychological characteristic not easily verified by observable evidence. Although clients' expressed evaluations can sometimes be partially verified by observations of clients' behavior (Bush and Gordon, 1978, p. 768 [3]), a client's expressed evaluation or satisfaction reflects an internal psychological state of the client that only the client can know directly.

Much valuable information that evaluators might obtain through client interviews is objective, not subjective. For example, the Health Interview Survey, a continuing national sample survey conducted by the U.S. Census Bureau, collects data on illness, injuries and other health topics. Since other people besides the respondent (i.e. a family member) might also have directly observed whether the respondent suffered from a particular illness or injury, these data are intrinsically objective, even for a respondent for whom no one is available to verify the information, and even for a respondent who supplies false information. Thus, objective data are not errorless data. Both objective data and subjective data can suffer from lack of reliability as well as validity. However, subjective measures involve special problems of interpretation and analysis in program evaluation, as will be discussed later.

Crime victimization surveys illustrate well the potential value of client interviews for obtaining objective data useful to program evaluation. Schneider (1976, pp. 136–137 [5]) reports that after an anti-burglary program was implemented in Portland, official crime statistics showed an increase in the burglary rate. In contrast, crime victimization surveys conducted before and after the implementation of the program revealed a decrease in the burglary rate, but combined with an increased willingness of victims to report burglaries to the police, resulting in an erroneous increase of the official burglary rate. Thus, without the victimization surveys evaluators may incorrectly have concluded that an effective program was ineffective.

Objective data obtained through client interviews, like data from agency records, also contain measurement error. Whether client interviews or agency records provide more reliable or valid objective data depends on the measurement errors inherent in each type. Client interviews are often done for only a sample of clients, resulting in sampling error and loss of reliability. Although agency records may be complete in the sense that all interactions with clients are recorded, the agency may have contact with only a subset of the potential clients, which may threaten the validity of the data for making inferences to the larger client population. In the case of crime statistics, victimization surveys suffer

from sampling error, whereas official records suffer from under-reporting. Thus, the choice between client interviews and agency records as a source of objective data about a total client population may sometimes involve a trade-off between reliability and validity.

Of course, client-derived objective data also can suffer from threats to validity. For example, Schneider (1975, pp. 13–14 [6]) hypothesizes that poor interviewing technique will cause crime victimization surveys to underestimate victimization rates and overestimate the proportion of crimes reported. To minimize these threats to validity, the interviewer must probe to make the respondent remember minor crimes that are more likely to go unreported. On the other hand, Levine (1976, pp. 311–325 [7]) argues that crime victimization surveys may overestimate crime rates because of lying, interviewer biases, memory failures about when crimes occurred, coding unreliability and mistaken interpretations of non-criminal incidents as crimes.

Since crime victimization surveys typically yield estimated crime rates 1.5 to 5 times larger than official rates (National Advisory Commission on Criminal Justice Standards and Goals, 1973, p. 199 [8]), serious biases must exist in one or both measures. This discrepancy is typically ascribed entirely to under-reporting biases in official crime rates. If a program analyst is willing to assume that the discrepancy between the survey results and the agency records is due largely to biases in the official records, then the analyst can compare victimization survey rates to official rates to compute ratios for estimating true crime rates from reported rates. Expensive victimization surveys need only be done occasionally, to ensure that the ratios have not changed greatly, and crime rates based on agency records can be inflated by the calculated ratios to estimate true rates. As this example shows, objective data derived from client interviews can usefully augment objective data from official records.

Intrinsically important vs not intrinsically important subjective measures

An often over-looked but critical distinction is between subjective measures that are intrinsically important, vs measures that are important only because of a presumed relationship with other variables. For example, contrast citizens' fear of crime with citizens' evaluative ratings of police services. Fear of crime and consequent feelings of insecurity decrease people's sense of well-being and enjoyment of life. In contrast, citizens' expressed evaluations of police services have no compelling *a priori* meaning; rather, interpretation must proceed from assumptions about the determinants of the expressed evaluations. For example, if the expressed evaluations reflect experience with and perceptions of services the police have provided, then the evaluations are a subjective measure of police performance. On the other hand, if expressed evaluations primarily reflect the feelings toward governmental authority found among different demographic groups, then the expressed evaluations are a measure of evaluative predisposition towards government. Clients' expressed evaluations of governmental programs differ therefore, from intrinsically important subjective measures of psychological states that contribute directly to people's quality of life, such as fear of crime, feelings of insecurity, general sense of well-being, sense of personal competence and feelings of stress and anxiety.

Since clients' expressed evaluations of governmental services have no clear *a priori* meaning or intrinsic importance, their appropriate use in program evaluation is debatable. Stipak [9] views client evaluations as performance measures only if they reflect some characteristics of the services actually provided, and emphasizes the need to establish a linkage between subjective evaluations and objective service characteristics before using client evaluations for measuring performance. Another perspective (e.g. Shin [10]) accepts on the basis of face validity that client evaluations and expressed satisfaction measure some aspect of the quality of service actually provided.

Both perspectives concerning the interpretation and use of client evaluations have disadvantages. The first perspective (Stipak [9]) suffers from the problem that client evaluations may be responses to subtle aspects of objective service performance not easily identified or measured. That perspective therefore biases program evaluators

towards discarding client evaluations as irrelevant. However, the second approach probably suffers from more severe problems. Public opinion research has found that citizens will readily express political opinions, despite knowing little about government or public affairs.† Similarly, clients may quite willingly provide evaluations of programs on the basis of little experience or knowledge. Client evaluations could therefore be meaningless, artificial creations of the interview process. Alternatively, expressed client evaluations could reflect real client attitudes that result from general feelings about government and public authority, from attitudes prevalent among client reference groups or subcultures, or from other causes irrelevant to the clients' experiences with the program being evaluated. The less contact the client has with the service agency, the more likely expressed evaluations will be meaningless. Thus, the conservative approach to interpreting client evaluations goes beyond face validity and examines the relationship of expressed evaluations to other performance measures and to service characteristics.

General vs specific client evaluations

Another important distinction concerns the specificity of the client's evaluation. For example, contrast the following two items from a client questionnaire used for monitoring mental health services (Urban Institute, 1978, p. A–3 [14]):

In an overall sense, how satisfied are you with the service you received?
Very satisfied, mostly satisfied, indifferent or mildly dissatisfied, quite satisfied.
Were the receptionist and clerical staff at the Center courteous and helpful?
Not at all, not much, somewhat, very much.

To cite another example, an Urban Institute report concerning municipal service evaluation recommends monitoring the percentage of households rating neighborhood park and recreation facilities as satisfactory, as well as specific evaluations concerning the condition of recreation equipment and the hours the facilities are available (Hatry *et al.*, 1977, p. 42 [15]). Thus, expressed client evaluations and satisfaction can range in specificity from evaluations of very explicit service characteristics to very general, global assessments.

Specific evaluations are probably less often meaningless creations of the interview process, since specific item referents are more likely to evoke responses based on the client's actual experience with the service agency or perceptions of service characteristics. Not surprisingly, specific subjective measures usually have higher reliability than global measures (Campbell *et al.*, 1976, p. 480 [1]). Skogan (1975, p. 58 [16]) criticizes the practical value of general evaluations for administrators, arguing that "general evaluations do not tell administrators what actions to take in the face of low ratings," and that "public officials need direct measures of those specific activities that are amenable to administrative manipulation." In a similar manner, Stipak (1979a, p. 51 [9]) argues that vague satisfaction and evaluation items confound in one indicator different aspects of service performance that should be measured separately. However, other researchers (Campbell *et al.*, 1976, p. 493 [1]) view global subjective measures as necessary expedients for requiring people to mentally summarize many specific factors. Evaluators can construct alternative general evaluation measures by combining a number of specific evaluation items, but this approach also suffers from problems (see Gutek, 1978, p. 52 [17]). Overall, evaluators can probably feel more confident of the reliability of specific client evaluations, compared to more general or global evaluations.

POSITIVE BIAS OF CLIENT EVALUATIONS

A problem in using clients to evaluate programs stems from the tendency of clients to provide highly favorable evaluations of all programs. Katz *et al.* (1975, p. 64 [18]) found

† Converse (1975, pp. 79–83 [11]) discusses information levels and opinion formation in the general public, and Stipak (1977, pp. 50–51 [12]) discusses processes of local political attitude formation in the absence of strong perceptions. For a discussion of meaningless responses to attitude items see Converse [13].

that about two-thirds of the clients of different governmental service agencies rated their satisfaction with the way the agency handled their problem as very satisfied or fairly well satisfied, compared to only about one-third answering somewhat dissatisfied or very dissatisfied. Across all services, 43% replied very satisfied, compared to only 14% replying very dissatisfied. As Campbell (1969, p. 426 [19]) has commented, voluntary and solicited testimonials from program participants provide an excellent source of favorable evaluations. Moreover, clients appear to express favorable evaluations and high levels of satisfaction regardless of program effectiveness. Scheirer (1978, p. 56 [20]) reviews a number of evaluation studies that found favorable client evaluations for programs that were not effective in achieving program goals. Perhaps the best example concerns the gastric freezing procedure for treating stomach ulcers (see Scheirer, 1978, pp. 53–54 [20]). By 1964, about 10,000 patients a year received this treatment and studies found that as many as 70% of the patients reported complete remission. Finally, in 1969 a large-scale, rigorous evaluation of the treatment was completed which showed that the treatment had no effect.

Other evidence also indicates there is a positive bias in people's subjective evaluations. Campbell *et al.* (1976, p. 99 [1]) found high levels of expressed satisfaction for a variety of different life domains. Only a small minority of the respondents admitted dissatisfaction or unhappiness, a finding Campbell *et al.* (1976, p. 99 [1]) point out agrees with the general finding from psychological research that subjects tend to use the positive side of rating scales more than the negative side. Another general research finding that shows a positive bias in people's subjective evaluations is the evidence (see Gutek, 1978, pp. 49–50 [17]) that people tend to evaluate more favorably their own lives and experiences, including experiences with governmental agencies, than the lives and experiences of other people. Similarly, Fowler (1974, pp. 149, 153 [21]) found that people think there is less crime in their own neighborhoods than other neighborhoods. In reviewing the extensive literature of job satisfaction, Taylor (1977, pp. 243–245 [22]) concludes that measured job satisfaction remains inexplicably high and that satisfaction levels do not vary with evidence of worker discontent such as absenteeism, strikes, or plant sabotage.

Why should subjective evaluations, in particular client evaluations of governmental programs, be so positively biased? Scheirer (1978, pp. 58–60 [20]) shows that social psychological theory would predict a positive evaluation bias due to several factors, including social desirability response bias, ingratiation attempts and cognitive consistency. Also, both program clients and program staff often receive a variety of benefits from participating in the program, regardless of the program's effectiveness in attaining official program goals (Scheirer, 1978, p. 59 [20]).

After reviewing the evidence and explanations for a positive bias of client evaluations, Scheirer (1978, pp. 61, 66 [20]) concludes that this bias is insidious in two ways. First, the processes creating positive evaluations are largely unconscious, and result from the client's social role. A positive bias therefore stems from more fundamental causes than gratitude, which Campbell [19] cites as the source for positive client testimonials, and will exist even when clients attempt to provide honest information for improving the program. Second, the positive evaluative bias helps to maintain ineffective programs, because of the demands and political pressure of program participants.

The positive bias of client evaluations necessitates a fundamental rule for program evaluation: Do not conclude a program is effective based only on the distribution of client responses on an evaluation or satisfaction rating scale.† A majority of clients will almost always choose a positive or satisfied response category. Evaluators should expect a majority of positive or satisfied responses and consider that an inconsequential finding. However, a finding of a majority of negative or dissatisfied responses indicates something unusual and provides a danger signal.

Some evidence indicates that specific evaluations exhibit less of a positive bias than

† However, evaluators can sometimes make inferences about program effectiveness by comparing responses for different groups, for different programs, and over time, as will be discussed later.

general evaluations. Campbell *et al.* (1976, p. 480 [1]) found less expressed satisfaction for the more specific and unambiguous life domains, compared to the more general and ambiguous. Bush and Gordon (1978, p. 777 [3]) found that clients of social welfare services provided consistently positive responses to general satisfaction questions, but were more discriminating in answering questions about specific aspects of the services. Thus, questions that refer to very specific program or service characteristics may tend to elicit a lower proportion of responses on the positive or satisfied side of the scale, compared to more general questions. Nonetheless, the evaluator does not know what biases exist for a particular evaluation or satisfaction item, or what interpretation to give different response categories. Conclusions about program effectiveness, based only on the distribution of client responses on *any* evaluation or satisfaction rating scale, are highly suspect.

CORRESPONDENCE BETWEEN CLIENT-DERIVED MEASURES AND OBJECTIVE CONDITIONS

The degree to which client-derived measures correspond to official records, to program characteristics and to the services actually provided has important implications for their use in program evaluation. This is true for client-derived objective measures, for intrinsically important subjective measures, and for client evaluations and expressed satisfaction. This section will discuss these implications, examine the findings of relevant research, and consider the importance of client expectations.

Implications for program evaluation

The implications of the correspondence between client-derived objective measures and official records was illustrated earlier, using the example of crime victimization studies. When there are no problems of validity, the choice between client-derived measures and agency records depends simply on cost and reliability—i.e. program analysts should collect those data that maximize reliability for a given cost. Thus, if crime rates estimated from victimization surveys and from official records were not biased relative to each other, program analysts should dispense with expensive victimization surveys. However, the huge discrepancies between official crime rates and victimization survey estimates reveal a problem of validity caused by under-reporting biases in official crime rates. As previously discussed, studies analyzing the correspondence between official rates and survey estimates can yield ratios for correcting the official rates. Thus, program analysts need not always resort to expensive victimization surveys to obtain valid estimates, as long as occasional studies monitor changes in the calculated ratios over time and across geographic regions. In this manner, careful analyses of the correspondence between client-derived objective measures and official records can alert evaluators to problems of validity, provide a method of correcting measurement biases and lower the costs of monitoring objective conditions and measuring performance.

The degree to which intrinsically important subjective measures correspond to objective performance and conditions determines the extent to which an agency's objective performance can change those subjective measures. Improving intrinsically important subjective measures, such as reducing fear of crime, is by definition a valuable public goal. Thus, research that establishes a correspondence between intrinsically important subjective measures and objective measures corroborates the importance of the objective measures. In fact, some objective measures, as Campbell *et al.* (1976, p. 3 [1]) argue, are important only because of an assumed relationship to the subjective experience of life. When studies fail to find a correspondence between objective agency activities and the intrinsically important subjective measures of concern, then changing those activities has no efficacy for improving those subjective measures. If no link to any normal agency activities can be identified, agencies should consider mass media campaigns and other efforts to affect directly people's subjective feelings. For example, Skogan (1977, p. 10 [23]) argues that the mass media's extensive crime coverage may heighten fear of

crime, which a public information campaign could reduce by presenting more realistic information on the crime problem.

The degree to which client evaluations and satisfaction correspond to objective conditions is critical. As discussed earlier, clients' expressed evaluations and satisfaction have no clear *a priori* meaning and intrinsic importance. Client evaluations can potentially be artificial creations of the interview process, or can reflect client attitudes that result from causes irrelevant to the program being evaluated. Whether client evaluations are valid measures of agency performance depends by definition on whether those evaluations do, in fact, correspond to any activities or services the agency performs. Therefore, research findings concerning the relationship between expressed client evaluations and objective measures are critical for interpreting the meaning of client evaluations, as well as for deciding on their appropriate role in program evaluation.

A potential complication concerning the relationship between client evaluations and objective conditions arises from possible simultaneity. Regardless of whether service performance affects clients' evaluations, clients' evaluations may affect clients' cooperation with program personnel, thereby affecting program performance. For example, Stipak (1979a, p. 52 [9]) hypothesizes that widespread dissatisfaction with police services may lower cooperation with law enforcement personnel. Variations in satisfaction across geographic areas, across demographic groups, or over time may consequently affect performance of police functions. Therefore, statistical attempts to estimate the impact of objective service conditions on subjective measures are potentially subject to simultaneity bias, resulting in spurious causal inferences.†

Although possible simultaneity complicates analyses that attempt to understand the causal relationships between subjective and objective measures, simultaneity does not complicate the practical task of measuring program performance. As long as a variable empirically corresponds to activities or services the agency performs, regardless of the direction of causal influence, that variable can serve as a valid performance indicator. Thus, the possibility that client evaluations may themselves affect objective program performance actually enhances their potential value as a performance indicator, even if actual performance has no effect on client evaluations.

A different type of simultaneity can potentially occur when both objective and subjective data are obtained from program clients. For example, analyses of crime victimization surveys (e.g. Parks [24]) typically interpret victims' evaluations of the police as a function of police response time and services rendered; however, Schneider *et al.* (1978, pp. 8–9 [25]) found that victims' attitudes toward the police affected victims' reports about response time and about police activities taken at the scene. This creates a serious problem for causal analysis, since measurement error in the objective variables is correlated with the subjective variable. Thus, analyses of relationships between client-derived objective and subjective measures are especially subject to simultaneity bias, resulting in spurious causal inferences. Moreover, this type of simultaneity, unlike the type previously discussed, does not enhance the value of the subjective indicator for measuring performance.‡

Client perceptions of service conditions

How accurately do clients perceive objective conditions and the actual services governments provide? Carroll [26] studied neighborhood residents' perceptions of their local streets, and found that about 80% of the residents correctly answered whether the street surface on their block was concrete or asphalt, and that over 90% correctly answered whether the street on their block had curbs. Schneider *et al.* [25] compared

† For example, assume an objective measure has no effect on a subjective measure, but that the subjective measure does affect the objective measure. Estimating a single equation model that regresses the subjective measure on the objective measure would reveal a spurious effect of the objective measure.

‡ Simultaneity in this case results solely because of the subjective measure's effect on measurement error in the objective variables. In the example previously discussed, simultaneity results because of the subjective variable's effect on actual objective performance.

objective data obtained from crime victimization surveys to official police records and found a close correspondence between the survey and police report data for factual details such as age and sex of suspects, number of suspects and events that occurred during the crime. In contrast, substantial differences were found for some factual details, including the race of the suspect, whether the victim knew the suspect, actions the police took at the scene and the month when the crime occurred. These discrepancies apparently resulted largely from errors in the information provided by survey respondents, but Schneider *et al.* (1978, pp. 5–6 [25]) were not able to determine the reason for such errors. A study conducted in Portland examined citizen perceptions of a large-scale program for improving and adding street lights in a target area. Only one-quarter of the citizens in the target area were aware of the new lights and almost two-thirds stated that no street lights had been added or improved (Schneider, 1976, p. 147 [5]). Finally, in the Kansas City Preventive Patrol Experiment the intensity of routine preventive police patrol was varied widely across areas within the city, but no effect was found on citizens' perceptions of the time police spent on patrol (Kelling *et al.*, 1976, p. 637 [27]).

As these research findings illustrate, client perceptions and client-derived factual information vary in how closely they correspond to objective conditions and actual service characteristics. Based on Carroll's [26] finding that citizens accurately perceive aspects of their neighborhood streets, Ostrom (1975, p. 9 [28]) concludes that citizens probably perceive fairly accurately specific attributes of other services. However, the accuracy of citizen perceptions almost certainly decreases rapidly as the objective conditions or service characteristics become less specific, less tangible and more removed from the citizens' immediate environment. Residents who accurately perceive salient physical characteristics of services on their block (e.g. type of street surface or presence of curbs) may have only vague impressions of less tangible neighborhood or community services. Clients of human service programs, social-welfare programs and other programs involving close interaction between clients and agency personnel probably have more accurate perceptions of at least some service characteristics than clients of programs involving less intense interaction between the agency and the clients.

Relationship between client evaluations and objective performance measures

A number of research studies have found almost no correspondence between subjective measures, such as client evaluations, and a variety of objective measures. Stipak's [9, 29] results showed little relationship between evaluations of local services in the Los Angeles area and different outcome, workload and input measures. According to the results of the Kansas City Preventive Patrol Experiment, large differences in the intensity of preventive police patrol have little effect on citizens' satisfaction with the police, on a variety of other attitudes toward the police, or on citizens' fear of crime (Kelling *et al.*, 1976, pp. 631–637 [27]). Several studies have found little or no relationship between attitudes toward the police and being the victim of a crime (Ostrom *et al.*, 1973, pp. 40–41 [30]; Smith and Hawkins, 1973, p. 140 [31]; McIntyre, 1967, p. 37 [32]). Analogous research on workers' job satisfaction has found little relationship between the level of satisfaction workers express and other measures of worker discontent such as strikes, absenteeism and plant sabotage (Taylor, 1977, p. 243 [22]).

Other research has found some correspondence, although weak, between subjective evaluations and objective measures. A program in Washington, D.C., to improve cleanliness in target neighborhoods apparently resulted in somewhat more favorable citizen evaluations of local street cleaning and alley cleaning services (Office of Policy Development and Research, 1978, pp. VI58–VI59 [33]). Schneider (1976, p. 148 [5]) found a modest relationship, within specific sub-areas in the Portland metropolitan area, between citizen evaluations of street lighting and interviewer counts of the number of street lights within sight of the respondent's household. Schuman and Grunberg (1972, pp. 376–377 [34]) examined city-level correlations between mean satisfaction with city services and objective measures such as the number of police *per capita*, the number of parks *per capita*, and the crime rate and concluded that "dissatisfaction is related at least

slightly to measurable aspects of city services." Finally, Marans and Wellman (1977, pp. 89–95 [35]) compared objective measures of water quality in northern Michigan lakes to subjective evaluations of residents living along the lakes. The subjective evaluations of water quality corresponded only modestly to the objective indicators of water quality.

Yet other research has found strong relationships between subjective evaluations and objective measures. Pelissero (1978, p. 36 [36]) reported that city residents living near a park provided more positive evaluations of local park and playground facilities, compared to respondents not living near a park. According to Skogan (1977, p. 7 [23]) people who have not been victims of crimes report less fear of crime than victims. Parks (1976, pp. 97–98 [24]) found that crime victims who reported high levels of satisfaction with how the police handled their call also provided more favorable general evaluations of police services and infers that how favorably victims respond to actions taken by police at the scene affects victims' general evaluations. Fowler (1974, pp. 59, 61 [21]) rank-ordered ten cities according to property tax rates and according to the percentage of city respondents who said local taxes were too high and concluded that the similarity in the two rank-orderings is "a very good testimony to the degree to which reality is reflected in people's attitudes." Finally, Aberbach and Walker (1970, p. 530 [37]) found that black respondents from Detroit who reported having experienced more racial discrimination gave more negative evaluations of city services.

Methodological flaws are one reason why different research studies have reached different conclusions about the correspondence between subjective evaluations and objective measures. For example, Peliserro [36] concludes a strong relationship exists between citizen evaluations of city parks and whether the citizen lives near a park; in contrast, Stipak [9, 29] found no independent relationship between distance to the nearest park and citizen evaluations of parks. Pelissero's analysis, however, does not include other independent variables (e.g. demographic variables) having possible confounding effects, whereas Stipak's analysis does. Even more seriously, Pelissero's objective measure of the availability of park facilities is the respondent's own report about whether a park is nearby. This measure almost certainly suffers from the problem discussed earlier of simultaneity between client-derived objective and subjective measures. In contrast, Stipak uses a very accurate, independently obtained measure of distance to the nearest park.

Spurious inferences due to simultaneity probably account for discrepancies between the conclusions reached by other studies, besides the Pelissero and Stipak analyses. Parks [24], for example, takes issue with prior research that concludes there is little independent relationship between crime victimization and evaluations of the police. Based on the finding that victims who were highly satisfied with how the police handled their call provided more favorable general evaluations of the police, Parks infers that actions of police on the scene improve victims' evaluations and thereby suppress the overall relationship between victimization and evaluations. But simultaneity may make this inference completely spurious, since initially having more positive evaluations of the police may predispose victims to report more satisfaction with actions police take on the scene.

Spurious conclusions that strong relationships exist between subjective evaluations and objective measures can result not only from simultaneity, but also from omitted variables, measurement error and data aggregation. The problem of omitted variables mentioned earlier when comparing the Pelissero [36] and Stipak [9, 29] analyses and the possibility of non-random measurement error in client-derived objective measures, also discussed previously, provide examples of how omitted variables and measurement error may create spurious subjective–objective relationships. Aggregating survey data to the level of cities or other geographic units maximizes the likelihood of spurious inference. Schuman and Grunberg [34] and Fowler [21], for example, aggregate citizen evaluations to the city level by computing city means and examine bivariate relationships between those means and city-level objective measures. This approach fails to statistically control for other variables that may have important effects on subjective evaluations, such as demographic characteristics of the respondent. The correct approach keeps the individ-

ual respondent as the unit of analysis and appends to the individual-level data objective data from other sources, as more detailed explications of these technical issues describe.†

Another approach to studying the impact of objective performance on subjective evaluations is the so-called "most similar systems" methodology some researchers have used to compare alternative institutional arrangements for providing local services (e.g. Ostrom *et al.* [30]; Rogers and Lipsey [38]). This approach is little more than what Campbell and Stanley (1963, p. 12 [39]) call the static group comparison design, except that some effort is made to choose groups for comparison that differ as little as possible on characteristics other than the treatment variable. However, since choosing two groups that match on all characteristics deemed relevant will usually be impossible, analysts should estimate the independent group effect by regressing the individual's subjective evaluation on a dummy group variable and other relevant independent variables (e.g. demographic variables), and not just compare frequency distributions or means for the two groups. Even then, measurement error in the independent variables and possible threats to internal validity due to selection‡ can produce overestimates of the independent effect that results from the objective inter-group performance difference.

Considering all the methodological problems that can spuriously increase the strength of relationship between subjective and objective measures, the overall results of existing research show amazingly little correspondence between client evaluations and objective service performance. A possible interpretation is that past studies have used inappropriate objective measures and that strong relationships do exist between service characteristics that clients care about and client evaluations. Further research concerning relationships between subjective and objective measures should attempt to identify, if possible, objective service characteristics that do affect client evaluations. As of now, existing research suggests that client evaluations usually do not correspond strongly to actual program characteristics.

Another important and largely unanswered research question concerns what factors affect the strength of the relationship between client evaluations and actual program characteristics. Campbell *et al.* (1976, pp. 478–482 [1]) hypothesize that the degree of correspondence between objective conditions and subjective judgements increases with the specificity and clarity of the object of judgement. Similarly, the previous discussion of client perceptions hypothesized that clients perceive less accurately those services that are less specific, less tangible and more removed from the clients' immediate environment. Lacking strong perceptions of such services, clients may express evaluations less related to actual service characteristics. Thus, expressed client evaluations of human service and social welfare programs may correspond more closely to program characteristics than would, for example, citizen evaluations of municipal police services or park facilities. Also, evaluations expressed by citizens who frequently use an available service, such as municipal park and recreation facilities, probably correspond more closely to objective service characteristics than do the evaluations expressed by nonusers.

Effect of client expectations

The expectations clients have for service performance probably decrease the correspondence between client evaluations and objective conditions. Assume that a client expresses an evaluation based on both perceived program performance and on performance expectations or standards the client uses for comparison. In that case, an improvement in perceived performance will produce a more positive evaluation. However, people may accommodate themselves over time to objective conditions by adjusting their aspir-

† For a comprehensive discussion of these statistical problems and their solutions see Hensler and Stipak [40]. For an analysis of when researchers can compute unbiased estimates of individual-level parameters from aggregate data see Firebaugh [41].

‡ As Skogan (1975, p. 49 [16]) comments regarding Rogers and Lipsey's [38] study of the effect of small versus large police departments, the citizens in the small, independent police jurisdiction had chosen by referendum not to be serviced by the large department. Skogan (1975, p. 49 [16]) argues that "all we know is that people who were so pleased with their local services that they voted to keep them are still pleased, while those who were not, are not."

ations and expectations (Campbell *et al.*, 1976, p. 485 [1]). Thus, an improvement in perceived performance may produce a more positive evaluation in the short term, followed by an increase in performance expectations, resulting in more negative evaluations. Similarly, a decrease in perceived performance may initially produce more negative evaluations, followed by decreased expectations and improved evaluations. Therefore, if client expectations adjust in a lagged manner to program performance, client evaluations may respond to short-term changes in objective performance, but adjustments of clients' expectations will ensure that only weak long-term or cross sectional relationships exist between client evaluations and objective performance measures.

Adaptation theory based on psychophysical experiments supports this view of the role expectations have in determining objective–subjective relationships. Although physical stimuli that depart greatly from a person's adaptation level excite a strong subjective reaction, repeated stimuli change the person's adaptation level and consequently lessen the subjective reaction (Helson, 1964, p. 227 [41]). Adaptation of expectations to objective conditions may explain Campbell *et al.*'s (1976, p. 466 [1]) finding that older blacks expressed higher satisfaction with their lives than either younger blacks or older whites, despite their impoverished material circumstances. Similarly, adaptation of expectations may partially explain the finding that the job satisfaction of workers in the same types of positions generally increases with job tenure (Taylor, 1977, p. 248 [22]).

The more that clients' expectations adjust to actual performance, the more the cross sectional and long-term relationships between client evaluations and objective performance measures will be attenuated. Thus, adjustment of expectations can invalidate using client evaluations to compare programs, or to monitor the same program over time. Evaluators making such comparisons should at least recognize the possible bias towards minimizing inter-program and over-time differences. Also, very high or low initial expectations of clients may greatly depress or inflate client evaluations of new programs, before clients' expectations have had time to adjust. Gutek (1978, p. 54 [17]), for example, speculates that widespread pre-existing negative views of public service agencies may greatly inflate reported satisfaction of new clients, since clients usually receive better service than they initially expect. Thus, the potential importance of client expectations in mediating objective–subjective relationships not only complicates the comparison of client evaluations across programs and over time, but also provides another reason, in addition to the positive bias of client evaluations discussed previously, for not using the distribution of client responses on an evaluation or satisfaction rating scale to assess the effectiveness of a program.

INTERPRETING CLIENT EVALUATIONS AS A BENEFIT OR PERFORMANCE SCALE

Client evaluations could be extremely useful to evaluation research if program evaluators could interpret them as a performance measure that scales programs according to program effectiveness, i.e. according to how much clients benefit from the program. In that case, evaluators could use client evaluations for comparing the performance of different programs, as well as for comparing the performance of one program over time and across different geographic areas or demographic groups. Unfortunately, using client evaluations as a performance scale can be erroneous for numerous reasons and is reasonable only under quite restrictive assumptions.

The previous sections have already covered some of the reasons why client evaluations cannot be interpreted *a priori* as a performance scale. Client awareness of service conditions and program characteristics appears to vary greatly. Similarly, existing research has not found a close correspondence between client evaluations and the activities and services agencies perform. Also, lagged adjustment of client expectations to objective conditions may weaken long-term and cross sectional relationships between client evaluations and actual program performance.

For these reasons, the first step towards a more sophisticated interpretation of client

evaluations requires recognition of the naivete of the simple consumer perspective that views client evaluations as the final judgment about program performance. Rating scales that require program clients to express an evaluation or a level of satisfaction do not necessarily measure the degree to which the actual program satisfies clients' consumption preferences. Even when expressed evaluations are linked to program performance, they probably reflect only some performance dimensions, and those only imperfectly. For example, Mechanic (1972, pp. 296–297 [42]) concludes that measures of patients' satisfaction with health services primarily reflect the personality and demeanor of the attending physician, not the physician's medical skill or the technical quality of the medical care. Patients' expressed satisfaction therefore measures limited aspects of program performance that may be unrelated to other performance aspects that the patients themselves probably consider critically important. Moreover, the physician's personality and demeanor are probably the aspects of performance least under the control of program administrators. As this example illustrates, program evaluators must recognize that client evaluations may at best tap only some aspects of program effectiveness and not necessarily those aspects most critical to program goals or most valuable for administrators to monitor.

The second step towards a more sophisticated interpretation of client evaluations requires an appreciation of the different possible types of client evaluation processes. Assuming that expressed evaluations are linked to some dimension of actual performance, they can result from several alternative processes that link expressed evaluations to perceptions of actual performance. Expressed evaluations may result from comparing the perceived effectiveness of some aspect of the program to a standard. If the standard of comparison is the ideal, best performance, which actual performance can only approach, then the client's expressed evaluation will increase monotonically with perceived effectiveness. Similarly, if the standard is an expected level, perhaps based on the client's past experience, which the client considers desirable not only to achieve but also to exceed, the client's evaluation will also increase monotonically with effectiveness. But if the standard is a specific, desired level of service, which the client considers desirable to achieve but not to exceed, the client's evaluation will not increase monotonically. For example, a specific, desired level of service could result from a rational consumer's calculation of the marginal cost and benefit of improved performance. For some clients, therefore, expressed evaluations may begin to decrease after some point with further increases in performance.

As Stipak (1979b, p. 424 [43]) points out, some clients may have complicated preferences that yield not only non-monotonic but also non-single-peaked evaluation functions, the functions that map the perceived performance dimension into the measured subjective responses. For example, some inner-city residents might desire a high level of preventive police patrol in order to minimize crime. At the same time, they might prefer a low level of patrol to an intermediate level, if they believed intermediate levels provoke more disruption than they prevent. Non-single-peakedness of the functional relationship between subjective measures and objective performance dimensions clearly increases the difficulty of analyzing subjective measures in order to provide administrators with useful information for modifying program performance.

The third step towards a more sophisticated interpretation of client evaluations requires an understanding of the issues involved in using client evaluations to compare performance for different individuals and groups. As already discussed, some evaluation processes may result in a non-monotonic relationship between evaluations and perceived effectiveness, at the level of the individual client. However, even if evaluations at the individual level are monotonic and can therefore be interpreted as a benefit scale, for several reasons monotonicity across individuals is not assured (Stipak, 1979b, p. 424 [43]). First, different individuals may base their subjective assessments on different aspects of service performance. Second, different individuals may apply different expectations or standards in evaluating program performance. Thus, even if each individual client will express more favorable evaluations the better the perceived performance,

clients expressing more favorable evaluations do not necessarily perceive better performance.

These issues concerning cross-individual comparisons of client evaluations are analogous to issues concerning utility in microeconomic theory. Whereas in microeconomic theory a consumer's utility function ranks the desirability of alternative consumption decisions, a client's expressed evaluation ranks the benefit of possible performance options, assuming a monotonic evaluation function as discussed previously. However, just as in microeconomic theory interpersonal comparisons of utility cannot be made in a theoretically sound way, cross-individual comparisons of client evaluations encounter problems of different expectations, standards, perceptions and evaluation processes.

Groups of clients that express higher average evaluations do not necessarily perceive better performance, due to all of the threats to individual-level and cross-individual monotonicity discussed above. In addition, variation in actual performance within each group complicates what sensible interpretations program evaluators can make from group averages or distributions on an evaluation item (Stipak, 1979b, p. 425 [43]). If considerable variation on actual performance does exist within each group, then attempts to rank-order groups must assume that either the shape of the distribution of the relevant objective performance dimension is similar within each group, or that no overlap exists between groups on the level of objective performance. Otherwise, attempts to rank-order the groups and to infer from group averages that one group experiences higher actual performance, are meaningless because of inter-group overlap. Finally, comparisons of group means assumes the reasonableness of treating the evaluation item as an interval scale.†

Because of these complications in interpreting client evaluations as a performance scale, eventually research may completely eschew the use of clients' subjective evaluations to measure program performance. That would be a mistake. Despite these theoretical complications, evaluators should not discard a tool of some potential practical value for measuring performance. Rather, evaluators must recognize what assumptions are necessary in order to consider a subjective evaluation measure an increasing monotonic function of some actual performance dimension. First, any reasonable analysis of a subjective indicator for purposes of measuring performance requires the assumption that clients base their responses on the same (or on empirically related) aspects of service performance, which the clients perceive fairly accurately. Second, individual-level monotonicity is required. An assumption of cross-individual monotonicity, however, is not reasonable or theoretically defensible—fortunately, it is unnecessary. What is necessary is an assumption that cross-individual differences are not systematically related to the client groups being compared, but rather are random differences that do not distort the inter-group comparisons.

To better understand the problem that cross-individual differences poses for program evaluation, consider this problem a result of the lack of a defined scale for subjective evaluation items. Different individuals scale the response categories to the items differently, in terms of the level of objective performance to which each category corresponds. Thus, the observed subjective measure has an error component due to interpersonal differences in scaling. The critical requirement for program evaluation is that this error component be statistically independent of membership in the groups the evaluator is comparing. In that case, lack of cross-individual comparability in scaling has no systematic effect on the relative group averages. However, if different groups tend to have different expectations or standards for service performance, systematic effects on the group averages can distort the rank-order of the groups on the subjective measure.

PROCEDURES FOR ANALYZING CLIENT EVALUATIONS

An obvious use of client evaluations and other subjective measures is for comparing

† See Hensler and Stipak [44] for a discussion of methods of estimating interval scale values for survey item response categories. For almost all applications in program evaluation, however, it is reasonable to simply assign rank-order numbers and treat the evaluation item as an interval variable.

programs, for comparing different client groups and for monitoring a program over time. For example, a program administrator might compare groups of clients, who had participated in a human service program for different lengths of time, in terms of their distribution of responses to an item asking for an evaluation of the services they received. Similarly, a city official might compare levels of expressed satisfaction with a municipal service in different geographic areas within a city, or a program evaluator might compare average satisfaction levels for participants in two alternative types of job training programs. The purpose of such comparisons usually is to judge relative program performance, not merely to describe expressed satisfaction. Whenever evaluators interpret client responses to an evaluation or satisfaction item as a reaction to actual program characteristics or agency services, that measure assumes the status of a performance measure, not just a measure of an internal psychological state of the client. Public officials may even proceed to reallocate program expenditures, based on comparisons between clients from different demographic groups or geographic areas (Webb and Hatry, 1973, pp. 20–22 [46]). As the previous section discussed, such comparisons require (1) client responses based on the same or empirically related performance aspects, (2) individual-level monotonicity, and (3) independence between cross-individual differences and group membership. This section will discuss analytical procedures that allow relaxing the last requirement.

Whenever cross-individual differences in scaling are related to the groups being compared, those differences will systematically distort comparisons of the group averages and distributions on the subjective measure.† For example, one group may have a disproportionate number of clients with especially high expectations for service performance and report lower satisfaction even though actual performance is higher. In order to make valid inferences in such cases about relative service performance therefore, evaluators cannot simply compare group means or distributions.

Rather than comparing group means, evaluators should use multiple regression analysis, often referred to as analysis of covariance (ANCOVA) in the evaluation research literature.‡ This technique can take into account other differences in the groups' composition, preventing those differences from distorting the comparisons of relative program performance. The basic approach involves using the client as the unit of analysis, and regressing the subjective performance measure on dummy variables distinguishing between the client groups being compared, plus variables for all other individual-level characteristics that might distort the inter-group comparisons. Stipak (1979b, pp. 430–432 [43]) presents simple simulation examples that illustrate how this method avoids erroneous conclusions arrived at by comparing group means. Although this method can potentially suffer from measurement error and other problems, as discussed later, it provides a far superior general method than direct comparison of group means for comparing client groups on an evaluation or satisfaction measure.

Evaluators using multiple regression analysis to compare client groups on a subjective measure should carefully consider what individual-level variables they need to take into account. Ideally, all differences in group composition that may create artificial differences between the groups must be accounted for by variables included in the regression equation. For example, if people of different socio-demographic characteristics tend to differ in their performance standards, in their susceptibility to some type of response set,§ or in other ways affecting the subjective measure, the evaluator should include

† Stipak (1979a, p. 50 [9]) demonstrates this in a more formal manner.

‡ Analysis of covariance is simply multiple regression analysis that includes both nominal-level and interval-level predictors. In the terminology of analysis of covariance a dummy variable, such as a variable indicating membership in a particular client group, is called a factor and an interval-level independent variable is called a covariate.

§ Campbell *et al.* (1976, p. 106 [1]) observe that comparisons of groups on a subjective indicator may lead to erroneous conclusions about group differences if group members differ in their susceptibility to a response set. For example, assume that people in one demographic group are more prone to acquiescence response set and that they perceive positive evaluations of a program as the socially desirable response. Members of that group will consequently tend to provide more positive evaluations, *ceteris paribus*.

variables representing those socio-demographic characteristics in the model. The evaluator can represent nominal-level client characteristics, such as race or sex, by using dummy variables. Since evidence exists that demographic characteristics such as race and age are related to client evaluations (e.g. Katz *et al.*, 1975, pp. 78–79 [18]), the conservative analytic strategy is to include variables for any socio-demographic characteristics by which the groups differ. If client evaluations were obtained prior to clients' intimate exposure to the program being evaluated, including that measure as an independent variable may help to account for differences in initial evaluative orientations.

Although far better in general than the simple comparison of group means, multiple regression analysis does not prevent all possible distortions of relative differences between client groups. Statistical biases can result from a number of possible causes.† Explicit measures of clients' performance expectations and general evaluative tendencies are usually not available. Analysts must therefore resort to using demographic variables as proxy variables, on the assumption that those variables correlate highly with the unmeasured variables. The weaker the correlation between the proxy variables and the unmeasured variables, the greater will be the distortions in the results of the regression analysis. Analysts sometimes introduce needless measurement error by categorizing continuous variables—an unnecessary practice in regression analysis, in contrast to crosstabular (contingency table) analysis.‡ By attempting to include relevant individual-level variables and to measure those variables as accurately as possible, program evaluators can minimize distortions of the estimates, obtained through regression analysis, of the relative inter-group differences that result from differences in actual program performance.

In most cases evaluators can probably reduce bias due to measurement error and omitted variables sufficiently to justify interpreting the estimated inter-group differences as possible reflections of differences in program performance. However, some potential problems can completely invalidate such interpretations. First, if important individual-level client differences are perfectly related to the groups being compared, no statistical method can separate the group-level performance differences from the effects of the individual-level variables. For example, if all clients in one program are of a different race than clients in another program, the differences on the subjective measure due to race and to the program are perfectly confounded. The smaller the client groups, the greater must be the heterogeneity within the groups to ensure the same level of accuracy. Similarly, individual-level client differences can also be confounded with differences in actual service performance. For example, program personnel may provide better service to clients of one race than another. When evaluators use geographic service areas to group clients for analysis, differences in service performance may become confounded with other differences across geographic areas. For example, citizens living in neighborhoods with dilapidated housing may tend to generalize their dissatisfaction with housing conditions to all aspects of their local area, including local governmental services (Stipak, 1979a, p. 49 [9]). These and other problems that seriously confound effects due to program performance with effects due to other variables make it impossible to extract any performance information from the subjective measure.

SUMMARY AND RECOMMENDATIONS

Client surveys can provide valuable objective and subjective information for monitoring and evaluating public programs. Crime victimization surveys and personal health surveys illustrate well the potential for client surveys to provide otherwise unavailable objective information and to correct validity problems in objective measures based on official agency records. When client-derived objective measures correspond closely to

† See Hensler and Stipak [40] for a concise overview of causes of statistical bias in such analyses and for greater detail see discussions in the econometrics literature of specification error and measurement error.

‡ Crosstabular analysis tends to produce not only measurement error due to categorization, but also specification error due to omission of relevant explanatory variables, since a small number of variables can be included at one time. Therefore, program evaluators should usually avoid crosstabulation.

official records, evaluators should simply use the combination of data collection techniques that maximizes reliability for a given cost.

Client surveys can provide information about intrinsically important subjective measures, such as citizens' fear of crime. When an intrinsically important subjective measure is causally linked to a public program or governmental service, administrators can potentially use that measure as one measure of performance† and can attempt to improve that measure through changes in program operations. When no causal linkage exists, officials may have no other recourse than public information campaigns for improving intrinsically important subjective measures.

Client surveys can also obtain information on clients' subjective evaluations of the program. The sensible use of clients' subjective evaluations and expressed satisfaction for purposes of program monitoring and evaluation involves a number of considerations evaluators must keep in mind. First, program clients are strongly biased towards providing highly favorable evaluations and expressing high levels of satisfaction. Second, the accuracy of clients' perceptions decreases for programs not involving close interaction with the agency and for less tangible governmental services that are removed from the citizens' immediate environment. Third, specific evaluation items usually have higher reliability and validity than general assessments. Fourth, even if actual program performance does not affect a subjective evaluation measure, that measure may reflect citizen attitudes that may themselves affect program performance. Fifth, using client evaluations to compare performance for different client groups requires client responses based on perceptions of the same or empirically related performance aspects, monotonic client evaluation functions and either independence between cross-individual differences and group membership, or else the use of statistical techniques that remove the distortions from those differences.

The sensible use of subjective data for measuring performance also requires the availability of necessary staff and data-processing capabilities. The analysis staff should definitely have training in multivariate statistics, especially multiple regression analysis, and preferably have some background in attitude measurement and scaling as well. The importance of statistical skills is probably greater for analyzing subjective measures of performance than for analyzing client-derived objective data or data about citizen preferences. The analytical staff will require access to some data-processing support or facilities; however, the widespread availability of statistical packages and the increasing use of computers and innovation in computing technology insure that data-processing will seldom present major problems. Almost any large public agency today can easily provide the necessary staff and data-processing support. Only very small public agencies may lack staff with the required statistical training or lack access to data-processing services. Those agencies should either forgo using subjective data for measuring performance, or else obtain capable help from consultants.

Outside consultants, agency staff, or program evaluators who use subjective performance measures should write a non-technical executive summary of their findings for general distribution. However, the findings should be based on rigorous analysis, usually documented by a technical report, that observes the following general rules.

Rule 1: Do not base conclusions about program effectiveness only on the distribution of client responses on an evaluation or satisfaction rating scale. Expect a majority of positive or satisfied responses and consider that finding inconsequential. However, recognize a majority of negative or dissatisfied responses as an unusual result and perhaps a danger signal.

Rule 2: Be alert for especially high or especially low client expectations. Unusually high or low client expectations, perhaps based on past experience, can greatly inflate or depress the initial evaluations of new program clients and of clients of new programs.

Rule 3: Look out for factors that may create non-monotonic client evaluation func-

† Note that using intrinsically important subjective measures for evaluating program performance involves some of the same analytical complications discussed regarding subjective evaluations and expressed satisfaction.

tions. Widespread publicity about the high cost of a program may cause clients to consider relative program costs and benefits, and to provide poor evaluations despite high levels of perceived effectiveness. Sometimes clients may perceive non-monotonic relationships between desired outcomes and agency workloads.

Rule 4: Ask how actual performance varies within each of the client groups being compared. If the groups overlap considerably on objective performance, judgements about overall group differences may be meaningless.

Rule 5: Use multiple regression analysis, rather than directly comparing group means or frequency distributions, when the groups differ on demographic or other obvious individual-level characteristics. Measure these characteristics as accurately as possible and include them as independent variables. Use proxy variables for relevant unmeasured characteristics.

Rule 6: Recognize when confounded effects make it impossible to extract performance information from subjective measures. The effects of actual program performance may sometimes be inextricably confounded with the effects of client characteristics and other differences between client groups and service areas.

REFERENCES

1. Campbell A., Converse P. E. and Rodgers W. L. *The Quality of American Life.* Russell Sage, New York (1976).
2. Andrews F. M. and Withey S. B. *Social Indicators of Well-Being*, Plenum, New York (1976).
3. Bush M. and Gordon A. C. The Advantages of Client Involvement in Evaluation Research, *Evaluation Studies Review Annual*, T. D. Cook (Ed.), Sage, Beverly Hills (1978).
4. Nunnally J. C. The Study of Change in Evaluation Research: Principles Concerning Measurement, Experimental Design and Analysis, *Handbook of Evaluation Research*, E. L. Struening and M. Guttentag (Eds), Sage, Beverly Hills (1975).
5. Schneider A. L. Victimization Surveys and Criminal Justice System Evaluation, *Sample Surveys of the Victims of Crime*, W. G. Skogan (Ed.), Ballinger, Cambridge (1976).
6. Schneider A. L. *Measuring Change in the Crime Rate: Problems in the Use of Official Data and Victimization Survey Data.* Oregon Research Institute, Eugene, Oregon (1975).
7. Levine J. P. The Potential for Crime Overreporting in Criminal Victimization Surveys, *Criminology* **14**, 307–330 (1976).
8. National Advisory Commission on Criminal Justice Standards and Goals. *Criminal Justice System.* U.S. Government Printing Office, Washington (1973).
9. Stipak B. Citizen Satisfaction with Urban Services: Potential Misuse as a Performance Indicator, *Public Administration Review* **39**, 46–52 (1979). Reprinted in *Evaluation Studies Review Annual*, L. Sechrest (Ed.), Sage, Beverly Hills (1979).
10. Shin D. C. The Quality of Municipal Service: Concept, Measure and Results, *Social Indicators Research* **4**, 207–229 (1977).
11. Converse P. E. Public Opinion and Voting Behavior, *Handbook of Political Science*, F. I. Greenstein and N. W. Polsby (Eds), Addison-Wesley, Reading, MA (1975).
12. Stipak B. Attitudes and Belief Systems Concerning Urban Services, *Public Opinion Quarterly* **41**, 41–55 (1977).
13. Converse P. E. Attitudes and Non-Attitudes: Continuation of a Dialogue, *The Quantitative Analysis of Social Problems*, E. R. Tufte (Ed.), Addison-Wesley, Reading, MA (1970).
14. Urban Institute. *What Happens to the Clients? Monitoring the Outcomes of State and Local Mental Health Services, Interim Report.* Urban Institute, Washington (1978).
15. Hatry H. P., Blair L. H., Fisk D. M., Greiner J. H., Hall J. R. Jr. and Schaenman P. S. *How Effective are Your Community Services? Procedures for Monitoring the Effectiveness of Municipal Services.* Urban Institute, Washington (1977).
16. Skogan W. G. Public Policy and Public Evaluations of Criminal Justice System Performance, *Crime and Criminal Justice*, J. A. Gardiner and M. A. Mulkey (Eds), Lexington Books, Lexington, MA (1975).
17. Gutek B. A. Strategies for Studying Client Satisfaction, *Journal of Social Issues* **34**, 44–56 (1978).
18. Katz D., Gutek B. A., Kahn R. L. and Barton E. *Bureaucratic Encounters: A Pilot Study in the Evaluation of Government Services.* Institute for Social Research, Ann Arbor (1975).
19. Campbell D. T. Reforms as Experiments, *Am. Psychol.* **24**, 409–428 (1969).
20. Scheirer M. A. Program Participants' Positive Perceptions: Psychological Conflict of Interest in Program Evaluation, *Evaluation Quarterly* **2**, 53–70 (1978). Reprinted in *Evaluation Studies Review Annual*, L. Sechrest (Ed.), Sage, Beverly Hills (1979).
21. Fowler F. J. Jr. *Citizen Attitudes Toward Local Government, Services and Taxes*, Ballinger, Cambridge (1974).
22. Taylor J. C. Job Satisfaction and the Quality of Working Life; A Reassessment, *J. occup Psychol.* **50**, 243–251 (1977).
23. Skogan W. G. Public Policy and the Fear of Crime in Large American Cities, *Public Law and Public Policy*, J. A. Gardiner (Ed.), Praeger, New York, Praeger (1977).
24. Parks R. B. Police Response to Victimization: Effects on Citizen Attitudes and Perceptions, *Sample Surveys of the Victimes of Crime*, W. G. Skogan (Ed.), Ballinger, Cambridge (1976).

25. Schneider A. L., Griffith W. R., Sumi D. H. and Burcart J. M. *Portland Forward Records Check of Crime Victims*, U.S. Government Printing Office, Washington (1978).
26. Carroll S. An Analysis of the Relationship Between Citizen Perceptions and Unobtrusive Measures of Street Conditions. Research report No. 10, Measures of Municipal Services: Multi-Mode Approaches Project. Workshop in Political Theory and Policy Analysis, Department of Political Science, Indiana University.
27. Kelling G. L., Pate T., Dieckman D. and Brown C. E. The Kansas City Preventive Patrol Experiment: A Summary Report, *Evaluation Studies Review Annual*, G. V. Glass (Ed.), Sage, Beverly Hills (1976).
28. Ostrom E. Multi-Mode Measures: From Potholes to Police. Paper presented at the Conference on Productivity and Program Evaluation: Challenges for the Public Service, organized by the Midwest Intergovernmental Training Committee (1975).
29. Stipak B. *Citizen Evaluations of Urban Services as Performance Indicators in Local Policy Analysis.* Ph.D. Dissertation, University of California, Los Angeles (1976).
30. Ostrom E., Baugh W. H., Guarasei R., Parks R. B. and Whitaker G. P. *Community Organization and the Provision of Police Services.* Sage, Beverly Hills (1973)
31. Smith P. E. and R. O. Hawkins. Victimization, Types of Citizen-Police Contacts and Attitudes Toward the Police, *Law and Society Review* **8,** 135–152 (1973).
32. McIntyre J. Public Attitudes Toward Crime and Law Enforcement, *The Annals* **374,** 34–46 (1967).
33. Office of Policy Development and Research. *Improving Productivity in Washington, D.C. Neighborhoods: A Case Study.* Washington: Dept. of Housing and Urban Development (1978).
34. Schuman H. and Grunberg B. Dissatisfaction with City Services: Is Race an Important Factor?· *People and Politics in Urban Society*, H. Hahn (Ed.), Sage, Beverly Hills (1972).
35. Marans R. W. and Wellman J. D. *The Quality of Non-Metropolitan Living: Evaluations, Behaviors, and Expectations of Northern Michigan Residents.* Institute for Social Research, Ann Arbor (1977).
36. Pelissero J. P. *Citizen Evaluations of Community Services in Oklahoma.* Bureau of Government Research, University of Oklahoma (1978).
37. Aberbach J. S. and Walker J. L. The Attitudes of Blacks and Whites Toward City Services: Implications for Public Policy, *Financing the Metropolis: Public Policy in Urban Economics*, J. P. Crecine (Ed.), Sage, Beverly Hills (1970).
38. Rogers B. D. and McCurdy L. C. Metropolitan Reform: Citizen Evaluations of Performances in Nashville-Davidson County, Tennessee, *Publius* **4,** 19–34 (1974).
39. Campbell D. T. and Stanley J. C. *Experimental and Quasi-Experimental Designs for Research.* Rand McNally, Chicago (1963).
40. Hensler C. and Stipak B. Contextual Analysis: Problems of Statistical Inference and Their Solutions. Unpublished manuscript available from Brian Stipak, Institute of Public Administration, Pennsylvania State University, University Park, PA 16802 (1979).
41. Firebaugh G. A Rule for Inferring Individual-Level Relationships from Aggregate Data. *American Sociological Review* **43,** 557–572 (1978).
42. Helson H. *Adaptation-Level Theory: An Experimental and Systematic Approach to Behavior.* Harper, New York (1964).
43. Mechanic D. *Public Expectations and Health Care.* Wiley, New York (1972).
44. Stipak B. Are there Sensible Ways to Analyze and Use Subjective Indicators of Urban Service Quality?, *Social Indicators Research* **6,** 421–439 (1979).
45. Hensler C. and Stipak B. Estimating Interval Scale Values for Survey Item Response Categories, *American Journal of Political Science* **23,** 627–649 (1979).
46. Webb K. and H. P. Hatry. *Obtaining Citizen Feedback: The Application of Citizen Surveys to Local Governments*, Urban Institute, Washington (1973).