

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

