Evaluating Gross Program Outcomes

Carl Simpson
Western Washington University

Chapter 3 (pp. 133-227) in:
Evaluating Social Programs at the State and Local Level: The JTPA Evaluation Design Project
Ann Bonar Blalock, ed.
Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 1990
DOI: 10.17848/9780585243900.ch3

Copyright ©1990. W.E. Upjohn Institute for Employment Research. All rights reserved.
3
Evaluating Gross Program Outcomes
Carl Simpson
Sociology Department
Western Washington University

Some scientists react to the difficulties of establishing cause and effect by withdrawing into their shells and refusing to say that the relationships they find are anything more than correlations. . . But decisionmakers cannot avoid making judgment. . . The decision maker want to know what to change so that he can achieve the effect he wants.

Julian Simon and Paul Burstein
Basic Research Methods in Social Science

General Concepts and Methods in Gross Impact Evaluation

Managers of nearly all organizations — human services and private sector businesses alike—examine the outcomes of their efforts and compare their own achievements with those of other similar organizations. Descriptions of outcomes and comparisons across organizations can be valuable management tools, but can also be misleading. The gross impact evaluation model offers guidance for maximizing the usefulness of these management tools while avoiding errors commonly encountered in the attempt. As Blalock argues in chapter 1, we want to facilitate evaluations that remain closely tied to management information needs, but go beyond program monitoring, both in terms of goals and methods.

The Perspective Taken in Conducting Gross Impact Evaluations

Evaluation research is often viewed as remote from service delivery—as serving distant purposes or as serving no purpose. However, the analysis of data on services and outcomes can be a valuable management
134 Evaluating Social Problems
tool guiding program development efforts. In addition, with the advent
of computerized management systems, such analysis has become fea-
sible in the majority of human services organizations as well as in the for-
profit sector. Data systems put in place to facilitate record-keeping and
report generation can be extended efficiently to provide a basis for the
analysis of service quality and effectiveness. Perhaps these factors help
explain why a recent survey finds many service delivery areas (SDAs) in
the Job Training Partnership Act (JTPA) system expressing a special
interest in systematic self-analysis (Seattle-King County 1985).

Systematic descriptions of program outcomes can focus policy plan-
ing discussions. Further, analysis that compares the effects of program
alternatives can identify strong and weak areas of current services, in
terms of their impacts on outcomes. The ability to focus change efforts
on low performance areas and to identify high performance approaches
as models for planning can enable a continuous improvement of services.
It amounts to “technological advance” for service organizations, where
effective technology—knowing what transformations produce desired
outputs—has been difficult to develop. This capacity to direct change
intelligently not only improves program services, but also provides staff
with a sense of efficacy—the sense that they are able to affect the quality
of their own work. Staff burnout has been identified as an ongoing
problem in job training organizations (Franklin and Ripley 1984). One
partial solution is putting the tools for more effective management in the
hands of local staff.

Even where a foundation of previous research has been laid, analyses
by local or state service delivery systems is valuable. Management
decisions must be specific: shall we implement services this way or that,
place participants with this type of trainer or that, deliver this set of
services or that? The local context determines which alternatives are
available and also the relative effectiveness of each. Thus, to apply
national research findings to local settings, while the only recourse in the
absence of better information, is a less reliable management guide than
developing local knowledge. For example, Wilms (1980) found no
difference between public and proprietary vocational trainers in South-
ern California, while Simpson (1982) found public schools substantially
more effective than proprietary schools for CETA participants in Washington State. To make policy decisions in either area based on research in the other area would be unfortunate.

*Gross impact analysis* is the study of program outcomes among program participants only. It resembles net impact analysis in that it focuses on outcomes and their probable causes, but it differs from net impact analysis in that no comparison group of nonparticipants is involved. Gross impact analysis provides quantitatively reliable knowledge about program quality and effectiveness, with the goal of guiding program development. The method builds efficiently on already-existing data collections adding data elements as required, to improve validity at a reasonable cost. Postprogram follow-up surveys are used to measure program outcomes. Program qualities tested for their possible influence on outcome levels are measured using available management information systems (MIS) and, where applicable, by collecting additional data describing services delivered to specific individuals and forms of program implementation developed by various service providers.

**Questions Gross Impact Analysis Cannot Address**

Among the wide range of possible impacts of any human services program, only a smaller set can be addressed using the gross impact approach. Several important society-wide goals of employment-related programs are essentially impossible to study definitively, because legislation that improves the situation of some individuals may be creating or overlooking problems among other individuals. These impacts include (1) increasing national productivity, (2) reducing total national unemployment, (3) reducing average job turnover time, and (4) improving the skill level and, therefore, the flexibility of the overall labor force.

In addition, gross impact analysis cannot draw any conclusions concerning the types or degree of change caused by participation in a particular program. This question can only be addressed by net impact studies, which compare program participants with similar individuals who did not participate in the program (see Johnson and Stromsdorfer, chapter 2). Gross outcomes refer to total postprogram outcomes; net impacts estimate the proportion of total outcomes that may be uniquely
attributed to participation in the program intervention. This means that gross impact studies cannot estimate the extent to which participation in a program changes individuals, the cost-effectiveness of a program, the time it takes for participants to repay the cost of a program in taxes generated by program success, or the impact of the program on reducing other costs such as welfare supports.

**Questions Gross Impact Analysis Can Address**

There are two broad categories of analysis goals for which the gross impact approach is well-suited: (1) describing a broad range of program outcomes, and (2) estimating the unique impact on outcomes produced by alternative methods of delivering services. The statistical assumptions underlying these and the power of the conclusions which can be drawn from them are so different that the remainder of this chapter will refer to them separately, as gross outcomes analysis and differential impact analysis. The first term avoids the word impact as a reminder that no cause-and-effect impact can be estimated using descriptive analysis.

**The Description of Gross Outcomes**

The description of gross program outcomes does not allow the evaluator to infer causation—to assume that the program or some aspect of the program is responsible for the outcomes observed. This type of analysis is well-suited to describing a wide range of outcomes for participants, employers, or others, with results available in a relatively short time. Descriptive findings in themselves imply no success or failure, but can be evaluated against managers’ expectations. Descriptive data may also help establish reasonable new baseline expectations for outcomes not previously measured systematically. Descriptive gross outcomes can also be used as tools to identify problem areas that deserve more detailed analysis. Finally, the ability to measure relatively numerous and detailed outcomes provides a way to describe the range of program outcomes, and how programs are achieving their impacts. For example, job training program outcomes can include such issues as whether employment is training-related, whether fringe benefits are provided, whether promotions are likely, and whether employers are satisfied with the programs.
Gross outcomes may be described for one office, one organization with multiple offices, or a system including more than one service delivery organization. However, where more than one organization or office is described, caution must be exercised in interpreting any comparisons that are made. Outcome comparisons tend to be interpreted by research consumers as if caused by differences in program effectiveness—an error, given the limitations of descriptive analysis.

The careful description of program outcomes is often a necessary element of process evaluations (discussed by Grembowski in chapter 4 of this volume). Two foci of gross impact evaluations—broadening the scope of outcomes measured and exercising technical care during measurement—are therefore particularly valuable to process evaluations. By the same token, a process analysis can be extremely valuable in identifying the particular outcomes that are appropriate to measure as part of a gross impact evaluation for a given organization.

**Differential Impact Analysis**

The second type of analysis proposed here does involve the estimation of cause and effect. Differential impact analysis is a method for rigorously comparing program variants—alternative service strategies and alternative approaches to implementing service delivery. Different program services and implementation forms can be compared to assess whether, other things equal, one or more alternatives are more effective than others in producing desired outcomes. That is, the unique impact of each program variant can be estimated in comparison to all other program variants being used in a program during the analysis. Participants experiencing each program variant act as a comparison group for those experiencing other variants. This opens the way to a wide range of analysis questions that might be asked by managers. The specific questions depend on what program variations exist in a particular service delivery system, and on the areas in which managers are most interested in developing information. The term *program variants* is used here as shorthand to include all existing program alternatives in services assigned to individual clients and in forms of implementation found among service providers in the service delivery system.
Within the constraints of sample size, these same questions can be asked for particular populations of participants: which treatment modes are most effective for target group A, and which for group B? Comparisons can also be made among service providers. This means that states can improve the reliability and meaningfulness of comparisons made among local agencies, and can also identify especially valuable directions for program technical assistance efforts, as long as the influences of environmental conditions and other important differences among agencies are taken into account.

Differential impact analysis is especially well-suited to program development efforts because of its ability to identify program variants that influence the program's ability to produce desired outcomes within the limits of the particular clients it serves and within the range of treatments available to it. It cannot tell us whether a program is worth retaining. However, given that a program exists, it can point the way toward making it operate more effectively.

Differential impact analysis recognizes that the most serious threat to reliable comparisons among program variants is selection bias. Clients who select or are selected for different program variants often differ from each other in ways that influence probable postprogram success. The impact of such client background differences must be accounted for before the unique impact of program variants can be identified. Otherwise, estimates of program variants are biased. Although no set of measures could ever estimate all selection effects, it is possible for analysts who know a particular service delivery system well to construct measures that will prevent a considerable proportion of the bias that can be caused by unmeasured selection. Each additional investment in such preventative steps improves the validity of the research findings.

The second major threat to differential impact analysis is from confounded program variants—service strategies or program implementation modalities that completely overlap one another. If all clients assisted through a particular service provider are assigned services that differ in, for example, three ways from services given all other clients in a system, it is impossible to determine which of these three might account for a higher or lower success rate by that provider. The treatments overlap
so completely that they cannot be separated statistically; they are con-
founded with one another.

Similarly, any time one service delivery organization implements
program variants that are entirely unique, the analyst encounters at least
two factors that are confounded: whether participants experienced the
specific program variant in question, and whether they enrolled through
the specific organization in question. Thus, completely overlapping
implementation by more than one provider as well as completely unique
implementation by any provider constitute confounded program vari-
ants, which make fully accurate differential impact analysis impossible.
This means that the differential impact analysis of treatment variants
becomes possible only when the treatment system being studied includes
a sufficient number of service providers or service tracks that implement
services differently, but not completely differently, from each other.

The Types of Factors Measured During Gross Impact Evaluations

Both descriptive outcome analysis and differential impact analysis
require measures of program outcomes. These may be recorded by
service providers, reported by participants during follow-up surveys, or
reported by employers. In addition, differential impact analysis requires
the measurement of program variants and “control variables.” Program
variants may be measured at the individual level, indicating which
specific services each participant received, or at the service provider
level, indicating how programs are implemented for the average client.
Control variables include a wide range of selection factors that can bias
findings if excluded from the analysis.

An overview of the various types of measures that may be involved in
gross impact designs is provided in exhibit 3.1. Measures are grouped
according to the source of each measure, and the purpose each measure
serves in an analysis (outcome, program variant, control variable). The
intersection of each type of outcome variable with each type of program
variant shown in the exhibit indicates a major relationship studied using
this approach. In exhibit 3.1, these relationships are indicated by letters
A through N. Any given differential impact analysis involves the relation-
ship between one outcome variable and some number of test variables
(program variants) along with the control variables included to protect against selection bias or other sources of error. Within this structure, each test variable represents an "hypothesized" effect on the outcome in question.

Exhibit 3.1 also indicates the most likely sources for each type of measure, recognizing that measurement decisions depend on the nature of each service delivery system. Outcomes are measured through observation of participants or relevant others. Some outcomes are observed by service providers, while others require special data collection. Surveys of service providers are inexpensive because their numbers are small compared to participants; however, they can measure only implementation variants, not individual service variants. Participant service records kept by service providers are both more reliable and less expensive than measures included during follow-up interviews of former participants.

### Exhibit 3.1

**Nature and Sources of Measures That May Be Included in Gross Impact Analysis**

<table>
<thead>
<tr>
<th>Data Source</th>
<th>OUTCOMES FOR Participants</th>
<th>PROGRAM VARIANTS Implementation</th>
<th>Service treatment</th>
<th>Controls against bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey of service providers</td>
<td>F</td>
<td></td>
<td></td>
<td>J</td>
</tr>
<tr>
<td>Participant treatment kept by service providers</td>
<td>D</td>
<td>G*</td>
<td>H</td>
<td>K</td>
</tr>
<tr>
<td>Standard MIS files</td>
<td>A</td>
<td></td>
<td></td>
<td>L</td>
</tr>
<tr>
<td>Participant follow-up surveys</td>
<td>B</td>
<td></td>
<td></td>
<td>I**</td>
</tr>
<tr>
<td>Follow-up surveys of others, such as employers</td>
<td>C</td>
<td>E</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>Published data by locality</td>
<td></td>
<td></td>
<td></td>
<td>N</td>
</tr>
</tbody>
</table>

* Individual service treatment records may be aggregated to indicate typical agency patterns.

** Selected treatment variants can be measured through participant follow-up surveys, although these require retrospective reports.
Variables measured as controls against bias can come from many sources, depending on the specific design of the analysis. Service providers, participants, and others, such as employers or referral agencies, can provide valuable data concerning the selection process. Service providers also implement policies that affect selection bias. This does not mean the analyst should ask service providers for their interpretation of their own selection processes. Rather, it means that once the analyst has identified specific selection policies and practices likely to affect outcomes, these will be measured most validly through reports from those directly implementing or experiencing the process. In addition, MIS files and published demographic and labor market data report standard variables known to affect labor market success and are, therefore, necessary to include in differential impact analyses. One strength of the gross impact approach is its flexible ability to measure multiple indicators of selection into particular service treatments.

The Design of Gross Impact Evaluation Studies

Nearly all modern human services organizations have committed records of client characteristics at intake and services received to the electronic memory banks of micro- or minicomputers. The relatively ready access an analyst has to these MIS records lays the initial base for inexpensive, yet valid, gross impact evaluation studies. Client populations can be defined and samples drawn from these MIS records, with individuals identified for inclusion in a study at either entry or termination from the program. Once the population—all clients who participated in the particular programs to be analyzed—is defined, representative samples can be assured by using a variety of convenient random selection methods. The population should include clients enrolled throughout a calendar year if the program experiences seasonal fluctuations in outcomes, and should proportionately represent all geographic areas from which clients are drawn.

MIS records may be augmented to improve the power of a given study to evaluate specific aspects of program services. In addition, standard survey research methods can be employed to perform short-term follow-up data collection on program participants or other relevant actors such
as employers, treatment facilities, or school administrators. For longer-period follow-ups, the problem of noncompletions reduces the value of the survey approach. For programs with mobile clientele, establishing good locator information on clients will be critical to the success of follow-up survey efforts. The construction of surveys, drawing of samples, and conducting of survey interviews are tasks about which much is known, making guidance readily available (e.g., Fink 1985; Babbie 1989; Rossi, Wright, and Anderson 1983; Frey 1983; Dillman 1978). In addition, the availability of university-based survey research centers, as well as private research firms, makes expert guidance readily available in most areas.

Analysis of gross impact data often can be performed on the same computers used to store MIS data, using one of the readily available statistical analysis packages. Any package that calculates percentage distributions, \textit{chi square} with percentaged tables, analysis of variance tests comparing means, and multiple regression is adequate for the needs of nearly all gross impact evaluation efforts.

The Application of Gross Impact Evaluation Concepts and Methods to the Case Example: JTPA

The remainder of this chapter illustrates the gross impact analysis approach by offering methodological guidance tailored specifically to one program, the Job Training Partnership Act. Much of what is said here applies to any job training program, and with a little translation, to any human services program. Conceptualization, design, measurement, and analysis issues in gross impact evaluations, however, are illustrated within the vocabulary of JTPA services and the JTPA service provider system. This section begins with a discussion of the proper uses of descriptive outcome analysis, then moves on to differential impact analysis and the study of employer outcomes.

The Uses and Limits of Descriptive Gross Outcomes Analysis

Descriptive analysis takes its name from its goal of examining outcomes without making causal attributions. Descriptive patterns may be
Evaluating Gross Program Outcomes 143

reported on the basis of data covering all participants or estimated from a sample of participants. Where sampling is involved, proper procedure will generate unbiased estimates of patterns characterizing all participants as well as information on how accurate those estimates are.

Descriptive data are relatively easy to collect and report, but also easy to misinterpret. This discussion, therefore, takes two directions: (1) identifying ways to make gross impact description most useful to JTPA managers, and (2) identifying the major limits on its valid interpretation.

Avoiding Interpretations that Imply Causality

As the term description indicates, the primary limitation on descriptive data analysis is that it involves none of the research design or analysis techniques designed for explaining causal relationships. The major reason for this limitation is that descriptive analysis offers no comparisons. For example, if we learn that employers are highly satisfied, we cannot know whether the reason is the friendly service, the quality of participants placed with them, reimbursement they may have received, a general tendency to answer positively, or a variety of other possibilities. We can guess, but the research findings offer no guidance until comparisons are made—in this example, comparisons between employers training more- and less-qualified participants, with higher and lower reimbursement levels.

This limitation does not mean that managers must refrain from interpretation. We all interpret the world daily. It means that managers must not assume that the findings themselves imply a particular interpretation. Thus, the first of the following two statements a JTPA manager might make is flatly incorrect, while the second could be correct.

1. "We find employers to be highly satisfied with JTPA, proving that we are sending them the types of employees they want."
2. "We find employers to be highly satisfied with JTPA. In my opinion, this is true because we are sending them the types of employees they want."

Statement 2 avoids incorrect causal attributions, while also stating a possible interpretation that could be examined through further analysis. The value of the descriptive finding is that it identifies the facts the
manager may work with and may attempt to explain. We can learn how satisfied employers are. The limitation is that the findings do not themselves offer any causal explanation for the level observed.

The most common error in interpreting descriptive data on job training programs is to assume that outcomes described after the completion of a program are caused by the program. In the heat of political battle, I may say, “Look what our program has accomplished; we have 84 percent placement rates!” In so stating, I may be taking credit for upswings in the economy, for individuals who recovered from temporary unemployment, and for random change, as well as for cases where employment was produced by the program. Similarly, if I claim that one service provider is “better” (causes greater success) than another on the basis of descriptive findings, I err by assuming that the difference was produced by program services alone, which cannot be demonstrated using descriptive statistics. Such claims are problematic not only because they are subject to self-serving interpretations, but also because they often tend to mire program development efforts in the most obvious and least useful interpretations—some vague improvement in “management,” “cream-ing” efforts, economic shifts, and so on.

Broadening the Range of Outcomes Measured

The most basic way in which descriptive analysis can improve a program manager’s information base is by taking advantage of its flexibility to enlarge the range of outcomes measured. We are too often wedded to the narrow range of outcomes readily available from agency records or required by government performance standards. These measures should be included in any descriptive analysis, but the addition of further outcome measures offers considerable benefits: the identification of unrecognized areas of program quality or problems, and the expansion of managerial decisionmaking into quality-enhancement rather than program-compliance only.

A hierarchy of outcomes may be arranged according to the extent to which each is required for a meaningful analysis of JTPA. Some states or SDAs may wish to include only a minimal core set of measures, making the follow-up as brief and inexpensive as possible and limiting
Exhibit 3.2
Prioritized Participant Outcome Measures

1. Required postprogram performance standards
   - Employed during 13th week after termination?
   - Earnings during 13th week
   - Number of weeks worked during 13-week follow-up period.

2. Other core measures explicit in JTPA mandate
   - Employment, including:
     - Hours per week employed at follow-up
     - Pre- to postprogram change in hours per week and percent weeks employed.
   - Earnings, including:
     - Hourly wage rate at follow-up
     - Total earnings from termination to follow-up
     - Preprogram to postprogram change in wages and earnings.
   - Welfare dependency, including:
     - Whether receiving public assistance at follow-up
     - Monthly dollar amount of public assistance at follow-up
     - Total public assistance received between termination and follow-up
     - Preprogram to postprogram change in public assistance received.

3. Measures of skill transfer and utilization
   - Whether employment is in training-related field.
   - Proportion of the work utilizing skills from training
   - For employer-based interventions, retention with that employer.

4. Measures of job quality
   - Benefits (medical, retirement plans; paid vacations; sick leave).
   - Likelihood of layoffs or reduction in hours
   - Likelihood of promotion and/or raises.

5. Measures characterizing those not employed at follow-up
   - Why termination job was lost or left, if applicable.
   - Whether participant is seeking work, and if not, why not.

6. Subjective orientations of participants
   - Intention to make use of the JTPA intervention (career orientation).
   - Personal evaluation of JTPA program services
   - Personal evaluation of postprogram job.
   - Personal comparison of postprogram job with preprogram job.
their analysis options accordingly. Others may wish to mount a more comprehensive analysis, once the decision is made to expend initial setup costs. The marginal increase in cost from inclusion of all six types of measures discussed below is small, making it logical to measure all. Nevertheless, some measures add information without being necessary to the research effort. With this distinction in mind, exhibit 3.2 displays outcomes in six categories, from highest (1) to lowest (6) priority.

The most basic outcomes focus on the explicit JTPA mandate that JTPA be considered an investment in individual lives—an investment in human capital. As such, it should show returns in higher probability of employment, higher earnings, and lower dependence on public assistance. Three measures are required by the Department of Labor. Beyond those, the survey method allows various components of employment and earnings, such as hours worked and wage rate, to be specified.

Although measures indicating skill transfer and utilization are not explicated as outcomes in the legislation, they are clearly implied. They represent the most direct impact of training-based interventions, and are especially sensitive to program variants, making these outcomes particularly useful to managers who wish to develop their programs based on differential impact analysis. They are also particularly useful for descriptive analysis, because findings indicate, in and of themselves, the extent to which outcomes are being produced via the method presupposed for all training programs.

In addition to wages, various intangible benefits from employment and indirect forms of income, such as medical benefits, are important aspects of job quality. A prime indicator of probable long-range employment success is whether the overall quality of each job places it into the category sometimes characterized as the "primary labor market" or into the "secondary labor market" (Doeringer and Piore 1971; Vermeulen and Hudson-Wilson 1981). Primary labor market jobs are relatively stable, include gradually improving income and benefit levels, are usually full time, include the possibility of promotion, include fringe benefits and are, in general, the types of jobs that can reasonably become a career. Secondary labor market jobs seldom include benefits, possibility of promotion, or a system of pay increments, are often part time, and are subject to layoffs. Even where a short-term follow-up shows partici-
pants retaining employment, jobs in the secondary labor market represent a poor risk for long-term employment stability.

Finally, the lowest priority outcomes are measures of participants' satisfaction. They are lower priority than other measures because their meaning is less clear, they are less reliably measured than other outcomes, and they have been excluded from most job training legislation and evaluation studies. Nevertheless, they are much to be recommended. They can offer valuable information to JTPA program operators, and they are inexpensive to add to participant interviews. In particular, subjective indicators of job adjustment may be extremely important to measure in cases where the transition into the workforce is expected to be especially problematic, e.g., in the case of long-term welfare recipients. Job satisfaction has surprisingly little correspondence to earnings, but is considerably influenced by job quality and skill use, making this subjective measure useful for assessing quality of program placements.

**Asking Questions**

One important approach to both limits and potentials of descriptive analysis is to ask meaningful questions without demanding more complex comparisons than allowed. Some questions involve no interpretation; they simply seek baseline descriptive information. Other questions may be worded specifically enough that a descriptive answer will assist the analyst in developing or confirming explanations. The following questions illustrate this point:

1. Does it appear that program goals are being met? If I know roughly which levels of program outcomes are expected, measuring outcomes lets me know whether I am in condition red, yellow, or green. Descriptive levels do not tell me why outcomes are higher or lower than expected, or whether my program itself has much to do with producing those outcomes. However, they tell me whether I need to look for factors creating low outcomes, whether high or low outcome levels are concentrated in particular program activities, and whether my organization is in better shape with regard to some outcomes than others. That is, descriptions of outcome levels can let managers know whether to worry, and which program areas to worry about most actively.
For some postprogram outcomes—those mandated as postprogram performance standards—clear expectations will be established. When expectations are unclear, descriptive measures can help establish reasonable baseline expectations. These would constitute first approximations that might be improved upon in subsequent

2. Does any service provider appear worth learning more about? One particularly useful application of descriptive analysis as a first approximation is the comparison of SDAs or subcontracting service providers. Such comparisons should be interpreted with great care, since agency performance levels are influenced by factors over which program operators have no control, such as the local economy, or may result from policies not intended by the act, such as increasing performance rates by serving those with least need. Descriptive differences point out where further investigation might be most useful, helping to pose questions correctly rather than answering them.

3. Is there any apparent change over time? Descriptive outcome figures kept over time—each month for example—can be used to form a baseline series indicating stability or change in services provided and program outcomes. Such a “time series” can sometimes alert managers to unexpected changes. It can also provide a relatively inexpensive first approximation of the effects of major program changes made during the time series.

*Investigating Specific Propositions*

One major strategy of multivariate analysis is to test a particular interpretation by seeing whether competing explanations for the observed findings can be eliminated. This tactic is not available for descriptive analysis. However, the same general strategy may be followed by posing questions thoughtfully and specifically enough to logically reduce the range of findings that would be consistent with the particular explanation proposed.

There is little value in asking broad questions such as: does on-the-job training (OJT) produce more placements than classroom training (CT)? Too many different interpretations could reasonably explain either
positive or negative findings. However, specific propositions direct expectations to only a few findings. If the expected finding occurs, then we have greater faith in the correctness of the proposition guiding the analysis.³

For example, if I identify some JTPA program activities as skill training programs, I will expect that a disproportionate number of postprogram job placements will be in the skill area. I have no a priori way to set expected levels, but descriptive findings are nevertheless interpretable. If only 2 percent of workers in my area are cashiers, and only 6 percent of my CT participants have previous experience as cashiers, then a finding that 65 percent of employed graduates from my cashier training class are cashiers suggests that the program is working in the way I envisioned. This does not indicate how well the program works, only that my proposed explanation about the way it works is supported.

Another example involves the question: what accounts for nonretention of jobs held at JTPA termination? Several specific propositions are easy to imagine, each suggesting its own specific measures. For example, if JTPA participants lack the ability to learn complex skills, instances of nonretention should occur most often when the training or the job involved complex skills or where the participant’s preprogram skills were weakest. Similarly, employers should often report that the participant was unable to perform complex tasks. If these variables are measured along with others indicating alternative explanations, managers can assess which explanations account best for the patterns observed.

As a final example, one may argue that JTPA should move participants into primary labor market positions (Taggart 1981). One could examine the degree to which this occurs by measuring qualities of postprogram jobs that define the primary labor market, as shown in exhibit 3.2. Findings would not indicate the degree to which JTPA treatment caused the job quality mix observed, but they would recommend greater or lesser concern about program quality, depending on the number of jobs exhibiting the desired qualities.

Perceptions Held by Employers and Participants

Some questions are inherently descriptive. If I wonder what impor-
tance employers place on various qualities of individuals they hire, I can ask them to tell me. Although it is always possible for data on such perceptions to be limited by incorrect self-knowledge or by misleading responses, these perceptions are appropriately interpreted in their descriptive form. The same is true of participants’ job satisfaction or other participant perceptions in which JTPA managers may have interest. Similarly, employers’ satisfaction with JTPA and their perceptions of the costs and benefits of participating in OJT or Work Experience may be taken at face value, as long as one recognizes that the information indicates no more than perception, and that perceptions do not necessarily reflect program impact.

Illustrations of Informative Descriptive Analyses

Several illustrations of descriptive findings are reported here, all taken from one statewide study of CETA OJT conducted in Washington State (Simpson 1984a). The findings displayed in exhibit 3.3 illustrate that measuring training-related employment as well as overall employment
at follow-up helps avoid jumping to an erroneous conclusion. If only the percent employed or not employed at follow-up were displayed, we would observe the expected pattern: AFDC recipients are employed at a rate about 9 percent below the rate for non-AFDC participants. We might, therefore, be led to conclude that AFDC recipients are less job-ready or less personally stable, and therefore fail more often than others to retain their OJT jobs. A program manager might consider imposing more counseling on AFDC participants, offering more support services during OJT contracts, or placing fewer AFDC recipients in OJT, although it is widely known to produce the highest postprogram placement rates.

However, when the outcome is presented with a slight increase in specificity, these interpretations no longer appear logical. AFDC and non-AFDC participants retain their OJT jobs at equal rates. They also move to other, training-related jobs at equal rates. The entire differential in employment is produced by the fact that among the 60 percent of participants who did not remain with their OJT employers or in the field of their OJT positions, non-AFDC recipients were two-thirds more likely to find work outside the OJT field. Now the most likely interpretation is that AFDC recipients are much less able to locate jobs without the assistance of the job training program, as shown by their relative lack of success once they leave the positions into which they were leveraged by OJT subsidies. However, once the program assisted their job entry, they retained their positions as often as their more employable colleagues. This means that the OJT program is doing well at equalizing the chances of AFDC people in the short run, but it also means that the OJT program effects are not carrying over to later job search success.4

Exhibit 3.4 illustrates the value of an extremely basic analysis of follow-up data. A group of OJT participants is followed from the beginning of their OJT contracts through contract completion and nine months beyond. The percent of the initial group who remain employed with their initial OJT employer has been calculated at several points during the OJT contracts, and monthly after termination. The findings graphed in exhibit 3.4 fall into three segments, so clear as to be quite valuable in their descriptive state.

During the OJT contract, some gradual attrition occurs, so that on the
date of contract completion, only 83 percent of the original placements remain. Then, a full 30 percent of all OJT jobs are lost or left in the month following the termination of the contract. Following this stage, attrition once again becomes gradual, with another 21 percent of jobs being left over a nine-month period. When the findings graphed here are combined with an official "entered employment at termination" rate of 78 percent, we also learn that the great majority of jobs lost during the first month following the OJT contract were counted as program successes. Despite their simplicity, these findings speak unambiguously about the value of postprogram follow-ups and the way in which the OJT system at that time and place was working.

The final illustration of descriptive findings offered here comes from a small recent survey of OJT employers in one SDA. The figures here suffer a relatively large error margin since they are based on only 78 interviews. However, employer responses were so extreme that the small numbers cannot obscure the basic thrust. Exhibit 3.5 reports one set of employer perceptions relevant to an interpretation of employer cost or benefit from participating in the JTPA OJT program. Aside from the wage subsidy OJT employers receive, the greatest cost or benefit they receive from participation in OJT is the work produced by the participant. If the participant is less skilled, slower to learn, less productive, or less
tractable than normal non-OJT hires, the placement represents a cost. Indeed, the wage subsidy is supposed to offset such costs. For this reason, employers were asked to compare their OJT participants, at the point of their initial hire, with the typical non-OJT hire for the same position.

The results shown in exhibit 3.5 make clear that in this SDA, the great majority of employers perceive that they have benefited from hiring an OJT participant. Over 80 percent say their OJT hire is easier to supervise than other hires, and over 60 percent say the OJT hire is a more productive worker than non-OJT hires. The same pattern holds for all the specific ratings save one: 31 percent of employers report their OJT participant needed greater-than-average training, while only 11 percent say they needed less. Even here, in the area assumed by definition to represent a cost to OJT employers, the majority say OJT and non-OJT hires are identical.

This descriptive finding is chosen to illustrate both the value and the limits of descriptive findings. Program planners can feel assured of the

<table>
<thead>
<tr>
<th>Total supervision ease</th>
<th>Total productivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honest</td>
<td></td>
</tr>
<tr>
<td>Gets along with others</td>
<td></td>
</tr>
<tr>
<td>Enthusiastic on job</td>
<td></td>
</tr>
<tr>
<td>Works overtime</td>
<td></td>
</tr>
<tr>
<td>Reliable</td>
<td></td>
</tr>
<tr>
<td>Follows directions</td>
<td></td>
</tr>
<tr>
<td>Learns new skills</td>
<td></td>
</tr>
<tr>
<td>Works fast</td>
<td></td>
</tr>
<tr>
<td>Works despite problems</td>
<td></td>
</tr>
<tr>
<td>Works independently</td>
<td></td>
</tr>
<tr>
<td>Trained at hire</td>
<td></td>
</tr>
</tbody>
</table>

---

**Exhibit 3.5**

Percent of JTPA OJT Employers in One SDA Who Rate OJT Participants Above, Equal to, or Below Typical Non-OJT Hires

<table>
<thead>
<tr>
<th>Criterion</th>
<th>OJT Worse</th>
<th>Equal</th>
<th>OJT Better</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total supervision ease</td>
<td>10</td>
<td>20</td>
<td>70</td>
</tr>
<tr>
<td>Total productivity</td>
<td>30</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>Honest</td>
<td>50</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>Gets along with others</td>
<td>60</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>Enthusiastic on job</td>
<td>70</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>Works overtime</td>
<td>80</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Reliable</td>
<td>90</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Follows directions</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Learns new skills</td>
<td>60</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>Works fast</td>
<td>70</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>Works despite problems</td>
<td>80</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Works independently</td>
<td>90</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Trained at hire</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
satisfaction with which their OJT program is being greeted, can use these results in marketing OJT to other employers, and can rest assured that their assessment and assignment system is locating successful workers. Yet, they cannot know how much of this employer satisfaction stems from excellent matching of particular employer needs and particular participant strengths and weaknesses, how much from the general employment maturity, which is by policy required of all OJT participants, and how much from "over-selection" of the most employable individuals to be placed in OJT. Many policy implications depend on these alternatives, although some implications are clear in any event, such as the feasibility of reducing wage subsidy without inducing employer perceptions that OJT is too costly.

**Minimal Research Design for a Participant Follow-Up Analysis**

Whether analysis will remain descriptive or move on to differential impact, certain minimal research design requirements guide the proper collection of survey data. A large amount of literature is available on survey research methods (see especially Rossi, Wright, and Anderson 1983; Dillman 1978). In this chapter, a set of topics that must be addressed by any client follow-up survey is listed with only brief comments on advisable design decisions for JTPA. A more extensive treatment of the design issues facing JTPA evaluations is found in Simpson (1986).

**Identifying the Population to Be Analyzed**

The *first* step in designing either a descriptive outcomes analysis or a differential impact analysis is deciding which set of participants to include—that is, how to define the *population* under study, the population to which conclusions will be generalized. In JTPA, the first such decision involves which authorizing titles are to be included. This choice depends primarily on managers' goals for the analysis. In addition, attention must be paid to the comparability of measures across title and to whether the programs operated under different titles are comparable enough to be combined meaningfully.
The second decision regarding which population to study is whether the population should include the following: all of those found to be eligible for JTPA services, all of those who were enrolled in JTPA, all statuses at termination, or only participants who were employed at termination. Studying all eligibles would be required for a full analysis of selection into the program. However, for a descriptive analysis of gross outcomes or a differential impact analysis of the effects of program service or implementation variants, only participants receiving services need be included.

There is sometimes a temptation to reduce data collection costs by including in the study population only individuals who terminated with employment. However, there are several reasons why gross impact analysis designs should include all termination statuses in the population to be studied. The only outcome that can be measured with a population limited to participants who were employed at termination is "retention of the termination job." Estimates of other standard outcome measures—average wage, proportion employed, etc.—would be badly inflated by excluding the group least likely to be employed at follow-up—those unemployed at termination. On the other hand, to estimate this group's follow-up employment and earnings at zero (their status at termination) would underestimate program success by ignoring delayed employment. Further, including all termination statuses allows the evaluation to examine why some individuals gain less than others from the program, insures comparability with cost data, and guards against differences in service providers' methods of defining termination status.

**Deciding Whether Data Collection Should Be Longitudinal**

In order to analyze change in employment, earnings, life satisfaction, and the like, the same individuals must be measured before and after the program intervention. In some situations, program eligibility data partially satisfy this need by detailing work history as well as individual background characteristics. When these data constitute adequate preprogram measures, an evaluation may be "added on" efficiently by identifying a sample at the point of program termination and measuring postprogram outcomes parallel to the preprogram measures already available.
A wide range of preprogram data may be informative concerning client characteristics. The important issue regarding the analysis of change is that the preprogram measures be similar enough to postprogram measures to be comparable. If the range of program outcomes is widened, additional preprogram measures may be required. However, pre- and postprogram measures need not be identical, as discussed later in this chapter.

**Determining the Duration of the Study**

While it would be convenient to concentrate data collection into a few months, this shortcut endangers the validity of the research, introducing known biases and others less easy to identify. The population to be studied should, therefore, be defined to include *all enrollees or terminees throughout the full year*.

This definition prevents bias due to seasonal variations in the labor market. Similarly, in classroom training, some institutions tend to end courses during particular months, so that the proportion of terminees who are program completers vs. those who are dropouts varies monthly. Third, different service providers develop different policies concerning when to commit their funds, in total or for particular services, and how long to hold unsuccessful participants before terminating them when required to by the expiration of agency contract periods. All these factors produce seasonal differences in the likelihood of program success.

**Defining the Size of the Study Sample**

Once a population is selected for analysis, the question becomes how large a sample should be in order for calculations to estimate accurately the patterns within the entire participant population. Assuming a *representative* sample, the primary determinant of error margin (i.e., of the accuracy of conclusions) is the number of cases upon which estimates are based. A conclusion that one program variant has 10 percent higher retention rate than another means little if the margin of error for that estimate is 20 percent. Therefore, the first decision must be how many cases are needed in order to generate a level of error acceptable to those who will use the analysis results.
One essential reason that survey research has become such a widely used method is that the accuracy of estimates rises rapidly as we move from very small samples to samples of modest size; yet, samples of modest size are nearly as accurate as very large samples. This occurs because error decreases as a function of the square root of the sample size. More precisely, the estimated error associated with any measure depends on the *standard error* of that measure. For random samples, the standard error equals the standard deviation of the measure divided by the square root of the sample size (i.e., $se = sd + \sqrt{N}$). Thus, if an income measure has a standard deviation of $4,000, the standard error is about $800 with a sample of 25, $400 with a sample of 100, $200 with a sample of 400, and $100 with a sample of 1,600.

One can see the danger of relying on a very small sample. However, it is equally evident that the marginal improvement from each increase in sample size is quickly reduced as sample size becomes larger. In the example above, adding 375 to a sample of 25 reduces error by $600, from $800 to $200. However, another 1,200 cases would be required to trim a further $100 off the standard error.

This phenomenon explains why many state and local surveys with limited funds choose sample sizes in the range of 350-500. Efficiency (accuracy gained per increase in data collection cost) rises rapidly below that level, but more slowly afterward. In addition, when percentages (e.g., percent "yes") are reported, samples in this size range produce error margins at or under 5 percent, a round number and error margin typically satisfactory for most purposes. However, sample size decisions should always be made after analysis goals are clearly established. Planners will be well-advised to consult one of several thorough texts on sampling (e.g., Kish 1965; Sudman 1976) or to employ a sampling specialist in cases where sampling appears problematic or in order to determine the most cost-efficient sample.

One additional consideration must be included in planning sample size: not all members of the initial sample selected will be contacted. Therefore, the number of specific individuals selected for inclusion in the sample must be greater than the number of completed interviews desired. The number to be selected is calculated by dividing the desired number
of completed interviews by the planned survey completion rate. For example, to complete 400 interviews at a completion rate of 70 percent would require that 571 names be identified in the initial sample (400 ÷ 0.7).

**Stratifying Samples**

Populations are sometimes divided into subgroups, or strata, each of which is sampled separately. Strata may be sampled in proportion to their numbers in the population, or disproportionately. Although some believe that samples must be proportionately stratified to insure the proper number of members with various background characteristics, this belief is in error. Proper sampling procedures insure a representative sample. No reason exists to consider proportionate stratification in gross impact analysis.

There are several conditions under which disproportionately stratified samples are sometimes recommended (Sudman 1976). Only one of these is applicable to gross impact analysis, but that one reason is central to statewide differential impact analysis as well as to analyses of subcontractor performance within large SDAs. The need for disproportionate stratification of gross impact samples arises when the analyst’s emphasis is on comparing or reliably characterizing subpopulations rather than on characterizing the entire population of participants. This occurs in the case of JTPA postprogram performance standards, where welfare recipients are treated as a separate stratum. In addition, statewide analysis aimed toward comparisons among SDAs, or SDA-level analysis comparing service providers, should consider stratifying to insure reliable characterization of smaller units. For such comparisons, the operative issue is not total sample size, but sample size for each subunit being compared with any other.

**Identifying Members of the Sample**

The sample of participants who are interviewed must be representative of the population being studied; i.e., each element of the population must have an equal chance to be included in the sample. *None of the claims for sample efficiency or reliability holds when samples are not representative.* Sample selection procedures must guarantee equal probability of
inclusion, eliminating any purposeful or accidental selection. The only way to guarantee equal probability of inclusion is to select from the population into the sample at random. The classical approach is to select each individual from an ordered list, using a table of random numbers. Two, more convenient methods, however, are equally valid.

First, the last three digits of participants’ social security numbers are random with respect to any meaningful characteristic of individuals. Therefore, sample members may be selected by identifying a range of three-digit numbers that would produce the required sample size, and including all participants with numbers falling in that range. If, for example, 25 percent of the names are to be sampled, the lower end of the range is chosen at random, and the upper is set at 250 higher. The second method is systematic sampling based on a random start. If 25 percent of the population is to be sampled, a list of names, typically a computer file, is prepared. One of the first four is chosen at random, and then every fourth name is included in the sample.

Participants may be selected into the sample at either program entry or termination, as long as the full population of participants is available for the sample. From the practical research administration viewpoint, it is often preferable to identify the sample at termination. However, if the data collection plan requires the addition of individual treatment measures throughout the program, it would be most efficient to identify participants for inclusion in the analysis upon entry. At that point, they could be specially tagged for collection of individual treatment data and inclusion in the postprogram follow-up.

Establishing a Follow-Up Period for Measuring Postprogram Outcomes

In addition to outcomes measured at program termination, postprogram outcomes are especially valuable for assessing program quality. In the case of JTPA, a three-month follow-up is required, making that period the obvious choice for a first, and perhaps only, follow-up. The majority of any survey costs occurs before the first question is asked—recording locator information, identifying a sample, keeping records on that sample, tracking hard-to-locate former participants, hiring and
training interviewers, setting up interview phone banks, making multiple calls to locate the participant, and introducing the purpose of the call. Therefore, the most efficient design is to add questions to a pre-existing survey.

In addition, the issue of selecting a follow-up period should be examined on its own merits. The major question is whether three months is too short a follow-up lag period. Factors weigh on both sides of the question.

There is value in extending the follow-up time period. Recent studies testing how well various follow-up measures predict long-term net impact of JTPA find three-month follow-ups much stronger than termination data alone, six-month follow-ups stronger than those at three months, and nine-month periods stronger than those at six months (Geraci 1984; Zornitsky et al. 1985). While the gain from each additional delay is smaller than the one before, each does offer improved validity.

Costs are also involved in extending the follow-up period. Follow-up surveys are subject to serious sample attrition if the first or only interviews are conducted very long after termination. Since sample attrition introduces unknown biases, it is preferable to conduct shorter-term follow-ups and achieve higher completion rates. In addition, a three-month follow-up is minimally acceptable. That delay is long enough to allow the rate of employment after classroom training to stabilize. It is also long enough for OJT placement to stabilize after the postcontract drop-off, even where 30-day delayed performance payments may delay that drop-off. Three months is also long enough to eliminate most inconsistencies introduced by the tendencies of some service providers to make more extensive use than others of the “administrative hold for job search assistance” category following the program.

Given the costs and inefficiencies of long-term follow-up surveys, states or SDAs planning to do longer-term follow-up may wish to consider using unemployment insurance (UI) wage records if they are available. Once access to the UI data base is established, a one- or even two-year follow-up is as easy to perform as a six-month follow-up. (UI system use is detailed by Johnson and Stromsdorfer in chapter 2 of this volume.) One factor, however, limits the usefulness of these data as a
The gross impact measure: UI data cover only individuals who maintain residence within the state. In the net impact approach, movement out of state is assumed to be equivalent for treated and untreated groups. However, gross outcomes are measured only for participants, making movement out of state a serious problem. One cannot determine whether a record of zero earnings represents continuous unemployment or movement out of the state. Use of this approach is recommended only if a separate tracking effort—to estimate the proportion who moved out of the UI reporting area—is mounted for those individuals with zero UI income. That estimate could then be used to adjust estimated job retention rates.

**Choosing the Data Collection Method**

Gross impact analysis involves data collection through follow-up surveys of participants and employers. The rapid expansion of the survey research industry has been accompanied by a growing literature on how to conduct surveys, the strengths and weaknesses of surveys, and the relative strengths and weaknesses of in-person interviews, mail questionnaires, and phone interviews (e.g., Dillman 1978; Rossi, Wright, and Anderson 1983). Survey research technology will not be detailed here; it will suffice to make the following summary claims:

1. Correctly conducted, surveys have proven highly reliable.
2. Surveys suffer much less response bias than once feared, as long as the respondent believes the interviews are conducted by a neutral party,8 and interviewers converse in a natural style (Bradburn 1983).
3. Bias from nonresponse can be problematic, but can be guarded against by achieving relatively high response rates and either insuring relatively equal response rates from all key subgroups or statistically adjusting subgroup response rates during analysis.
4. Given the cost of in-person interviews and the high nonresponse rate and possible educational bias in response with mail questionnaires, phone surveys are usually recommended for program evaluation.

**Conclusions Concerning Design**

When the research design decisions just discussed have been made, the
basic structure of the evaluation effort is in place. These issues have been touched on only briefly, as they are so basic. However, detailed treatments are available in various standard texts (e.g., Rossi and Freeman 1982), making further comment here unnecessary.

**Differential Impact Analysis:**
*Estimating Influences on Program Outcomes*

The primary goal of differential impact analysis is to reliably describe differences in postprogram outcomes across program services and forms of service delivery implementation, so as to identify the *probable causes* of those differences. In nonexperimental research, identifying causal relationships is problematic. However, quasi-experimental research designs, such as those recommended by Campbell and Stanley (1966), can considerably increase our confidence in having reliably identified the major causes of differences we observe. (See also Campbell and Cook 1979; Caporaso and Roos 1973.) For purposes of program development, managers will wish to gain information about the differential effectiveness of *program variants*—options available to managers in the services assigned to individuals and in forms of program implementation. These variants may be altered on the basis of evaluation findings in order to improve program effectiveness.

Formal, quantitative differential impact analysis is a relatively underdeveloped field, as is well-illustrated in Borus’s (1979) program evaluation primer. After listing 44 specific participant characteristics known to affect labor market success, he turns to the question of “program component independent variables.” His one-paragraph discussion of this topic begins by saying that “It would be extremely useful in modifying existing programs and in the planning of new programs to know which of the (program) components is most effective for various types of participants” (Borus 1979, p. 70). Two measures are suggested: program length, “... and, if possible, a measure of quality.” A review of previous studies using a differential impact analysis approach to evaluating job training programs is included in Simpson (1986).

Identifying probable causal connections is valuable to program managers because changing a factor that has a causal influence on program
outcomes is likely to change the level of those outcomes. While formal causal modeling may seem the domain of esoteric social science, it is essential to research on the basis of which investment decisions may be made. Let us suppose that a weak analysis confirms higher success rates among individuals who receive shorter intake procedures and concludes that brief intake causes improved program performance. Let us further assume that a more solid causal analysis would have learned that the most highly employable participants entering the system were given brief intake because they had little need, and that it was their employability rather than the intake that affected their postprogram success. Ironically, if an SDA were to base decisions on such weak research and overhaul its intake system to offer only short intake, its performance would not improve and might deteriorate, because those participants who needed, and previously received, the longest intake no longer have that opportunity.

The goal of each of the steps involved in differential impact analysis is to increase our confidence that we have identified those program variants that do have a direct influence on program outcomes and are therefore useful to program managers in improving their programs. This chapter can only summarize some major characteristics of nonexperimental research designed to increase the analyst's ability to identify causal relationships. There are also useful references available on this complex subject (e.g., Blalock 1964; 1985).

Research into causal relationships begins with comparisons. To determine whether program option A is better than option B, one must identify a criterion of comparison (e.g., job retention) and compare options A and B on that dimension. Options could be basic program activities, such as OJT, classroom training, or work experience, or optional variants within the same activity, such as OJT assignments developed by the participant, developed by the service provider, or initiated by an employer. These comparisons should be selected so that a causal interpretation is reasonable. This is where past research findings, economic theory, and managers' knowledge of programs come into play. If answering a question in causal terms would fly in the face of logic or of established information, the question probably should not be posed as part of a differential impact analysis.
In addition, to convincingly establish that a relationship is causal, findings from our comparisons must hold up after competing explanations have been eliminated. Each time we identify a plausible competing explanation, test it, and find that it does not explain away the difference between options A and B, we increase our confidence in the causal association between program variant A/B and the outcome in question. For example, our confidence in finding that public classroom trainers perform better than proprietary trainers (or the reverse) would be increased by learning that the difference in performance could not be accounted for by differences in participants’ literacy skills or prior job experience, by differences in the fields for which each type of school trained participants, etc.

The goal of quasi-experimental research is to eliminate the effects of all important measurable alternative explanations. That goal is never reached, but we can eliminate many important alternative explanations. These include both factors of interest to the analyst, such as other program variants confounded with the one being tested, and control variables, such as age or gender, known to affect the outcome in question.

Classical experiments attempt to eliminate competing explanations by controlling variants other than the A/B comparison of interest and by randomly assigning individuals to variants A and B, hoping thereby to produce groups equivalent in all regards except for the variant under study. Quasi-experimental research occurs in settings that allow neither the control of variants other than those directly under study nor the random assignment of participants to program variants. Instead, multivariate statistical techniques are used to determine whether alternative explanations are able to undermine our confidence in findings.

The primary strategy of differential impact analysis is to utilize each program variant as a comparison group for each other variant. Except where program variants are too highly correlated with each other or with participant background characteristics, multivariate analysis can estimate the unique effects of each. This same approach has guided recent research on the impact of college (Astin 1977). In the case of college impact, no untreated comparison group exists, making net impact studies
impossible. However, comparisons among colleges, with each acting as comparison group for the other, are possible given careful measurement of differential selection into each college. The same logic applies to differential impact analysis, where the decision is made to structure research that cannot ask net impact questions but which can, nonetheless, reliably compare different treatments and treatment contexts.

**Combating Bias That Threatens the Validity of Differential Impact Analysis**

The comparisons demanded by the analysis goals and the data collection method—surveys in this case—determine the major threats to the validity of differential impact analysis. These risks are summarized below. Each is a source of bias, as opposed to random error. The term bias refers to error that consistently misdirects research results. Like a compass with a metal object nearby, readings from the analysis are distorted in a consistent direction. To correct the findings, one must remove the object or adjust for its influence. Random error differs from bias in that it takes no particular direction. Random error can be as serious as bias if it is large. However, techniques for minimizing random error are well-developed in survey research (i.e., careful measurement techniques and properly constituted samples).

Some types of bias can also be dealt with through standard survey research techniques. These include bias from censored samples, nonresponse bias, and response bias. They can be prevented by selecting samples correctly, achieving high response rates across all major groups in the sample, and wording questions properly. The two types of bias defined in exhibit 3.6, selection bias and bias from confounded program variants, are combated during multivariate analysis. This means that descriptive gross outcomes analysis, which does not employ multivariate techniques, is always subject to serious bias. Differential impact analysis is able to reduce, but not eliminate, these biases during analysis.

Bias is reduced when equations include measures indicating selectivity and program variants that overlap with the program variant being tested. However, many selection biases are unknown or cannot be measured, making statistical adjustments difficult. Therefore, selection
Selection Bias

When participants who select or are selected into different program variants differ in ways that affect program outcomes, observed outcome differences between program variants could be produced either by program qualities or by participant characteristics. If participant characteristics are not taken into account, estimates of program impact will be biased. For example, preprogram employment experience influences postprogram employability and may also influence which program service is assigned. If employment history is not measured, the analyst cannot adjust for its impact on postprogram outcomes, producing biased estimates of the differential impact of service assignment. Since many such differences may exist but not be measurable (e.g., subjective motivation), some degree of selection bias is always present in differential impact analysis.

Confounded Program Variants

When two program variants are correlated with each other, the unique effects of each can be estimated if both are included in the same equation. However, if one is omitted, then the estimate for the included variant will absorb the effect of the omitted one, biasing conclusions in that direction. For example, if in a particular service delivery system, service providers who exercise especially careful quality control over employer sites acceptable for OJT assignments also conduct more elaborate intake assessment of job maturity prior to OJT assignment, and if both of these improve OJT postprogram outcomes, then both must be included in the analysis, or the estimated impact of either one alone will be inflated.

Minimizing Bias from Nonrandom Selection

Reducing the effects of selection bias follows the general logic of causal analysis. Each source of selection bias is an alternative explanation that can be countered only by inclusion in multivariate equations of variables which identify the selection process. The discussion below identifies four major sources of selection bias. For each, the aspects of differential impact analysis most likely to be affected and the measurement strategies most able to minimize the bias are indicated.

Sources of Selection Bias

1. **Legally eligible individuals may or may not apply to JTPA** because of differences in information available, personality or motivational differences or geographical differences in services available. This selection process is critical for net impact studies, but seldom biases differential impact analysis, which involve only comparisons among individuals already enrolled in JTPA. However, if this type of bias
differs across SDAs, then statewide differential impact comparisons among SDAs will be affected.

Little protection from this type of bias is available to analysis that does not include an untreated comparison group. However, SDA-level measures of program availability and participant measures of motivation for applying to JTPA may help assess possible differences between SDAs. In addition, standard demographic background characteristics may be correlated with motivational characteristics, allowing their inclusion in differential impact analysis to act as a partial proxy for direct measures of motivation.

2. Participants may or may not be enrolled into JTPA after eligibility is determined. If the reasons are correlated with program outcomes, bias will result. The source of this type of selection may be program policies and practices such as targeting, a participant’s choice after learning of program options, or failure to locate a program placement of the type decided on for that participant.

Measuring the source and nature of the selection is the appropriate tool for reducing bias from these sources. It is possible to measure implementation policies such as targeting, which are intended to determine which eligible individuals are enrolled. In addition, measures of participants’ demographic and work history characteristics may act as a proxy for agency selection or may indicate which participants best fit the agency’s desired targets. Beyond that, the key measures of agency selection involve the proportion of eligibles for each provider who fail to enroll and the reasons why they have made that choice. Given some JTPA managers’ reported emphasis on enrolling the most qualified participants, statewide differential impact analysis will be well-advised to include agency-level measures of intended and, where possible, actual selectivity by service providers.

3. Participants may request a particular treatment aside from the decision to enroll in JTPA. If the reasons for that request also predict that participant’s likely program outcome, this self-selection can bias estimates of how program activity affects outcomes. This is especially likely to occur where employers select desirable job applicants and then send them to JTPA to request OJT enrollment,
or where schools send their best students to JTPA for support to complete a program. Participant requests are most likely to involve a basic program activity, or a particular school or employer. Analysis comparing these most basic treatment variants is, therefore, the most likely to suffer from this source of bias.

The best protection against bias from self-selected treatment is measuring participants’ route into JTPA: whether they requested particular services, and if so, which ones and why. In particular, the route from employer or school to JTPA should be identified, since it involves a type of self-selection most likely to affect postprogram success. One could also measure the degree to which particular service providers control the assignment to treatment vs. allowing participants to elect their own treatment.

4. Participants may also be assigned to particular treatments by program managers. If treatment A rather than treatment B is assigned on the basis of factors that also influence program outcome, selection bias is present.

This source of bias has potentially pervasive effects on differential impact analysis because rational service provider policy offers the most intensive services to those with the greatest need. That is, many JTPA services are intentionally compensatory. Since “greatest need” often translates to “least employable,” the selection of services on the basis of need can bias estimates of treatment impacts on employment outcomes in studies that do not include measures of need in the analysis. To identify compensatory effects of treatment, one must have measures of both the need and the treatment. Since these two factors have opposite effects, they cancel each other out and neither effect is visible without joint analysis of both variables.

Measurement Approaches to Combat Selection Bias

When differential selection cannot be prevented, it must be identified by measuring the selection process and decisions. Differential impact analysis has the advantage of attacking from two different angles, using individual-level measures of preprogram characteristics and of the selection process experienced by each individual participant, and also employing agency-level implementation measures of selection policies
and typical practices. Individual-level measures have several advantages for combating bias:

1. Ideally, they can include the agency's diagnosis of each participant's need and its service prescription for each participant as well as the treatment each individual actually received. Both the participant's true level of need and the agency's perception of each participant's need are important potential sources of selection bias.

2. They offer information on the explicit selection process by the JTPA agency. For example, we can learn how much intake time the agency spent with each particular participant. It is one thing to know what "full" intake includes (an agency-level measure) and another to know that individual A received only a "fast track" intake review, while individual B was judged to require extensive pre-employment services.

3. They can include the route each specific individual takes into training. This proves to be one primary indicator of selection bias.

4. They allow precise measurement of program variants, increasing the power of the measures most important to any differential impact analysis.

Agency-level variables also exhibit two particular strengths in combating selection bias.

1. The problem of compensatory treatment cannot be fully solved by individual-level measures, because no precise measures of need or assistance exist. Agency-level measures indicate resources available or provided on average. They are, therefore, much less influenced by compensatory treatment. For example, if agency A provides job search assistance to only 5 percent of clients while agency B does so for 40 percent, it is almost certain that many individuals of equal need will receive this service in agency B but not in agency A.

2. Agency policies directly affect selection. Targeting decisions, policy toward "creaming," policies regarding single vs. multiple activity treatments, and the like, have some consistent effect on selection across all participants enrolled through a particular agency. Such agency policies can indicate selection on difficult-to-measure criteria such as how participants present themselves interpersonally.
The Potential for Severe Bias in Analyses of the Most Basic Program Divisions

Perhaps the two most basic program divisions managers might wish to analyze using the differential impact analysis approach are different service providers within a service delivery system and different program activity assignments. Unhappily, these divisions are the most likely to be affected by selection bias as well as by bias from confounded program variants. Implementation policies vary most widely across service providers and across basic program services. Service providers are likely to want control over which participants are assigned to each basic service. Similarly, participants are more likely to exercise choice regarding preferred basic services than about more specific implementation policies. Service provider implementation and service mix involve basic resource allocation decisions and are, therefore, likely to be affected by geopolitical concerns.

Different Service Providers

SDAs, and to a smaller extent their subcontractors, are located in different labor markets and political atmospheres. Although one can account for some of these differences through measures of the labor market environment and agency policies, many will remain unmeasured. Therefore, some unknown degree of bias will persist in analysis across service providers, especially SDAs. The fuller and more accurate the measures of labor market environment and agency policies concerning recruitment and selection, the smaller the remaining bias.

Basic Program Activities

Basic service treatment options are designed, in part, to accommodate differences in participant needs and qualifications. In particular, job search assistance assumes job readiness, OJT assumes minimum acceptability to employers, and work experience assumes an absence of even the most basic job experience. Selection bias is likely to be especially serious in such cases, because differential selection on the basis of employability is explicitly called for. In addition, different treatments produce outcomes through different mechanisms, making
them complex to compare directly. For example, true training interventions are intended to produce skill transfer and employment, leading to a career line. These outcomes have little relevance, however, for an intervention based on securing employment through the leverage of a payment made to employers.

These concerns lead to the recommendation that (where resources allow) sample size should be large enough to accommodate separate analysis within each basic activity, along with tests performed across all activities. In addition, differential impact analysis should include “membership-identifier” variables indicating enrollment in each of the most common program activities and in each SDA included in an analysis. These variables will absorb the effects of basic program activity and also some unmeasured selection effects, thereby reducing bias in estimates of other effects.

Although these problems appear overwhelming, and are never solved completely, each additional measure of selection improves estimates. The analyst, therefore, has the power to produce highly useful and quite accurate estimates, which should, nevertheless, always be interpreted as imperfect. Simpson’s (1989) analysis of one SDA illustrates the value of measuring selection into program activities. Without controls, it appeared that enrolling participants in multiple-sequenced activities produced only a slight improvement in job retention at 13 weeks. However, after including in the analysis a set of competency benchmarks measured at program entry, a large benefit became observable for those enrolled in sequenced activities. These competencies had been used as a basis for assigning multiple services, so that participants with greatest need received most intensive treatment. These two factors tended to cancel each other out, so that the effects of each could be observed only when both were included in the equation.

**Measuring Potential Influences on Outcomes**

Differential impact analysis tests the impact of program variants and control variables on postprogram outcomes. In particular, managers may test whether particular forms of implementation produce greater or lesser success and whether assignment to particular program services or to
services with particular qualities improves an individual’s chances of experiencing postprogram success. Basic categories of factors may be tested to examine their possible impact on these outcomes. They are shown in exhibit 3.7.

As discussed earlier, the analysis of how program implementation and treatment influence outcomes is one of the most neglected areas of

<table>
<thead>
<tr>
<th>Types of Determinants</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Program variants: Forms of program implementation</strong></td>
<td></td>
</tr>
<tr>
<td>Program variants</td>
<td></td>
</tr>
<tr>
<td>Program variants</td>
<td></td>
</tr>
<tr>
<td>Individual services (treatment) received</td>
<td></td>
</tr>
<tr>
<td>Membership identifiers</td>
<td></td>
</tr>
<tr>
<td>Variable descriptions</td>
<td></td>
</tr>
<tr>
<td>Control variables</td>
<td></td>
</tr>
<tr>
<td>Client characteristics</td>
<td></td>
</tr>
<tr>
<td>Selection processes</td>
<td></td>
</tr>
<tr>
<td>Labor market characteristics</td>
<td></td>
</tr>
<tr>
<td>Basic organizational components</td>
<td>• Form of contracting used</td>
</tr>
<tr>
<td></td>
<td>• Program cost</td>
</tr>
<tr>
<td></td>
<td>• Size of program</td>
</tr>
<tr>
<td>Service delivery framework</td>
<td>• Targeting (selection) policies</td>
</tr>
<tr>
<td></td>
<td>• Typical intake procedures</td>
</tr>
<tr>
<td></td>
<td>• Quality control procedures</td>
</tr>
<tr>
<td></td>
<td>• Exit practices</td>
</tr>
<tr>
<td></td>
<td>• Basic service assigned to (OJT, CT, etc.)</td>
</tr>
<tr>
<td></td>
<td>• Service provider enrolled through</td>
</tr>
<tr>
<td></td>
<td>• Intake screening intensity or method</td>
</tr>
<tr>
<td></td>
<td>• Treatment intensity (length, complexity)</td>
</tr>
<tr>
<td></td>
<td>• Characteristics of the trainer</td>
</tr>
<tr>
<td></td>
<td>• Job search assistance received</td>
</tr>
<tr>
<td></td>
<td>• Age</td>
</tr>
<tr>
<td></td>
<td>• Employment history</td>
</tr>
<tr>
<td></td>
<td>• Educational attainment</td>
</tr>
<tr>
<td></td>
<td>• Indicators of &quot;creaming&quot;</td>
</tr>
<tr>
<td></td>
<td>• Referrals involving prescreening</td>
</tr>
<tr>
<td></td>
<td>• Unemployment rates across place/time</td>
</tr>
<tr>
<td></td>
<td>• Median wage across place/time</td>
</tr>
</tbody>
</table>
research in employment and training. Implementation studies do not include estimates of impact on outcomes. Outcome studies typically work with the very limited base of treatment measures derived from accessible agency records. This means the thoughtful measurement of program variants is the key to creative advances in program development based on differential impact analysis.

Of the factors shown in exhibit 3.7, those labeled "program variants" are identified by choice, because the analyst hopes to learn whether particular implementation forms or particular individual services enhance program outcomes. Program variants include the basic organizational arrangements and the service delivery framework established to implement delivery of services, and the specific treatment received by each participant. Treatment includes both *variable descriptions* of the services received, such as the length of a training program, and measures identifying only whether or not a participant was a member of some particular set, such as recipient of classroom training or enrollee through a particular service provider. A third set of factors, "control variables," is required in order to insure that effects estimated for program variants are as accurate as possible. These measure participant background, selection processes, labor market qualities, etc. When these are analyzed together, it becomes possible to isolate estimates of the unique effects of each on a given program outcome.

The conceptualization phase of measuring treatment and implementation is critical. The criteria summarized in exhibit 3.8 offer some guidance. Criteria 1 and 6 are essentially technical, and may be honored without much knowledge of JTPA. However, criteria 2 through 5 require knowledge of the state or local JTPA service delivery system, making input from program managers critical to successful analysis. *As a general principle, differential impact analysis becomes useful to guide program development only when measures are developed in collaboration between researchers and program directors.* Measures are also suggested by JTPA implementation studies (reviewed by Grembowski, chapter 2 in this volume) and some excellent analyses of CETA implementation (Levitan and Mangum 1981; Snedeker and Snedeker 1973; Franklin and Ripley 1984).
174 Evaluating Social Problems

Exhibit 3.8
Criteria Useful When Selecting Implementation
or Treatment Variables for Analysis

1. Can the variable be measured reliably?
2. Is there reason to believe it varies across individuals or service providers?
3. Is there reason to believe it represents a nontrivial program impact?
4. Does the variable measure a program variant under the control of program managers; that is, is it a policy-relevant variable?
5. Are program managers open to changing the program variant to be measured?
6. Given cost and time constraints, can the measure be integrated into a data collection scheme?

Service Provider Surveys to Measure Program Implementation Variants

In exhibit 3.7, the first sets of measures characterize program implementation—aspects of the organizations put in place to provide JTPA services. These measures characterize the entire organization—its structure, policies, and practices—rather than any one participant's treatment. Most program implementation variants are best measured through surveys of service-providing organizations. Data for each service provider can then be attached to the computer files of all participants who enrolled through that provider. In this way, data collected inexpensively by surveying a limited number of service providers can be used to analyze program impacts on all individual participants in the sample.

Depending on the nature of the measures, agency directors may be able to answer reliably, or agency staff may need to compare notes or consult records in order to characterize typical practices accurately. With easy-to-answer questions, phone surveys may be used. For more demanding measures, however, a written survey, which allows time for data gathering, is preferable. The recommendation here is to use a written survey of each service-providing organization, with a backup telephone contact person who can clarify questions as they arise.

One problem with service provider surveys is that agencies may intend one form of implementation but actually carry out another, making self-descriptions inaccurate. This can be partially remedied by a second form of implementation measurement: aggregated agency characteristics.
Aggregated variables are measured with the individual participant as unit and then summed, percentaged, or averaged across all participants within each agency. For example, agency surveys could report whether policies emphasize training women for nontraditional occupations. The aggregated form of measurement for this same issue begins by constructing the individual-level variable. Nontraditional training fields for women are identified, and the training field of each female participant in the sample is coded as traditional or nontraditional. That individual variable is then aggregated for each agency, producing an agency-level variable, "percent of female participants trained in nontraditional fields," which may be used to double-check agency reports.

Suggested Measures of Program Implementation

Program implementation variants can be divided into basic organizational components, such as forms of contracting and staffing, and service delivery framework within which intake, service assignment, provision of basic services and support services, and program exit occur. The latter are most likely to have a direct influence on program outcomes, because they affect the nature of services provided and the selection process through which individuals are assigned to treatments. Yet basic organizational components are inexpensive to measure and may influence outcomes indirectly, by affecting various aspects of the service delivery framework. They are also important as control variables, to protect interpretations from alternative explanations after the research is completed.

Basic Organizational Composition

Forms of contracting. SDAs may or may not use requests for proposals (RFPs) as part of the subcontracting process. For both service provider contracts and trainer referral contracts, use of fixed-price, performance-based contracts may be contrasted with other approaches to contracting.

Staff qualifications. Franklin and Ripley (1984) argue strongly that staff qualifications represent a key to success, although they are not specific about what constitutes good qualifications.

Staff turnover. One might assume that staff stability (low turnover)
would predict success, although during the late CETA era, one study found the reverse to be true (Simpson 1984a).

Staff workload and division of labor. Client-to-staff ratio, ongoing staff training, and the division of staff between direct service, administration, and development work can influence intensity and effectiveness of service delivery and, therefore, client outcomes.

Service provider history. The age of service providers, how much their services have changed over time, their relations with the private and public sectors, and their rate of growth or decline may be useful to identify, although JTPA implementation studies suggest these factors have little differential impact on program outcomes.

Size. The size of SDAs or subcontractors (amount of grant, number of participants, size of staff) may also be included as control variables.

Program costs and cost-related policies. Program costs are usually very difficult to measure precisely for individuals or for specific services. However, total cost-per-participant can be measured as an implementation variable, and analyzed either as outcome or as one possible influence on outcomes, when different SDAs or different service providers are being compared. In addition, policies toward use of support services, rate of employer reimbursement, and length of training can be measured. Aggregated measures, such as the proportion of participants receiving support services or the average length of OJTs, may also be useful.

Service Delivery Framework

Many of the same factors Grembowski poses as key to understanding organizational process (see chapter 4) are also valuable measures of program implementation for differential impact analysis. These may serve as program variants or as controls against bias.

Explicit selection processes. Agency selection is critical to measure, both as a service quality issue and also because selection bias can be partially addressed with such measures. Agency policy may emphasize enrolling the most job-ready, those with greatest need, or those whom the program is most likely to benefit. Agency selection policy may reserve some activities, such as short OJTs, primarily for those who are most easily served. A greater or smaller proportion of participants may have
been referred initially by employers or schools, adding an issue of preselection.

**Intake procedures.** Procedures used during intake for selection, diagnosis, information giving, and counseling may differ in intensity and type. They may affect how well the agency treatments match the abilities and needs of each participant to the labor market. They may also act as indicators of the agency selection process.

Possible measures include length and intensity of typical intake; proportion of participants who get full intake; number of “hurdles” participants must pass, as an indicator of selection for motivation; whether intake is centralized or conducted by subcontractors; whether intake is conducted individually or in groups; what diagnostic tools are available and how often various tools are used, policies regarding targeting and other screening criteria; and what proportion of placements with employers were initiated by the employers.

**Quality control over referral and program activity mix.** Service providers may exercise strict control over the development of participant assignments, including rigorous screening of schools, agencies, or employers involved in treatment, or they may take the *laissez faire* approach, offering information and encouraging participants’ self-directed search for assignments, but exercising little control. This issue promises to be one of the most valuable areas for agency-level measurement, because referral represents the pivotal point of agency influence over treatment.

Possible measures of agency control include whether employer referrals are encouraged, giving control to employers, or carefully reviewed and screened, retaining agency control; whether policy encourages “open contract” referral arrangements, whereby employers agree to fill all openings for certain job titles by choosing among a set of eligibles sent for review by JTPA, giving service providers great control; whether the agency conducts formal quality reviews of employer-trainers or classroom-trainers, increasing control; and whether a large proportion of assignments are developed through participants’ self-directed search, which decreases agency control.

The second half of the quality control issue is *how quality is defined.*
Among those agencies that perform explicit quality control reviews over potential referrals or placements, what criteria are used to define quality? Some possible measures include the following: previous JTPA placement or retention track record, if applicable; in the case of employer-based interventions, employer stability or growth, typical non-JTPA turnover rates, typical wage rates, amount and quality of training likely to occur in OJT assignments, employer's ability to supervise constructively; in the case of schools or community-based trainers, staff quality, ability to handle special student needs, placement assistance for graduates, and credibility among employers.

Exit practices. The final element of treatment is the set of program completion and job search options implemented. Exit practices are especially important to measure at the agency level. Individual measures of job search assistance suffer from the compensation problem: those least able to locate jobs on their own are most likely to receive job search assistance, creating the appearance that postprogram employment is negatively correlated with receipt of job search assistance. Agency-level measures of the average availability of assistance are not affected by the compensation problem. Measures may include availability of job clubs or job search workshops, proportion of staff time devoted to postprogram job development, proportion of trainers who include formal job placement assistance, and method of placement, i.e., centralized or handled by subcontractors.

Individual Measures of Treatment Variants

The nature and intensity of services received by different participants in a program vary widely within service providers as well as between them. These differences among individual experiences require individual level measurement. In addition, individual-level measures offer several advantages, as listed in exhibit 3.9.

Some individuals' treatment experiences are routinely recorded as part of agency MIS files. Others may be recorded by agency staff as the treatments occur, or may be included in participant follow-up surveys, measured through participants' recall of the services they received. The preferable form of measurement is agency recording. Agency staff can
Evaluating Gross Program Outcomes

Exhibit 3.9

Major Advantages of Individual Level Measures

- They tie program services to outcomes for the same specific individuals, offering precise analysis of the degree of association between the two.
- They tie specific services and outcomes to specific individual background characteristics, providing direct tests of control variables.
- Normally, they vary more widely than agency level measures, strengthening statistical tests.
- Normally, they suffer less overlap with other test variables than agency level measures, strengthening statistical tests.

record services as they occur, avoiding recall errors, and including sequence. Staff are also more able than participants to identify which services are administered. Agency measurement is also less expensive, as measured in terms of dollars, since it avoids telephone interview time. It is, however, more expensive in terms of staff time and often in terms of staff resistance to data recording.

Gathering such information through participant surveys also has certain advantages: less lead time is required; it is the only option available when samples are identified at termination; and state or multi-SDA analyses may not reflect information necessary to coordinate data collection by large numbers of local direct service personnel, making measurement at follow-up the only viable option.

Intake Services, Screening, and Selection

Since intake intensity should be compensatory, with greatest intensity reserved for those with greatest need, good intake will tend to equalize the chances for success of those with greater and lesser need. Accurately estimating the impact of intake on program outcomes, therefore, requires both measures of intake experiences and also measures of participants’ need for intake assistance. The best approach to measuring need is to identify specific barriers to employment via the competency-benchmarking approach developed most thoroughly in the area of youth competencies (Simpson 1989).

Measures of individual intake experiences can include both the nature and intensity of intake diagnosis and services. One approach would measure time in intake (individually or in workshops), specific intake diagnostics such as testing, and intensity of specific types of intake.
services such as career counseling. Another approach would identify separate paths taken by individuals enroute to program enrollment, such as employer-initiated OJT contracts.

**Delay Between Eligibility and Enrollment**

The time lag between eligibility and enrollment may be a component of selection bias. Enrollment that is almost immediate typically indicates a referral initiated by an employer or a school, and therefore involves preselection. Beyond that, up to some point, delays tend to weed out the least motivated. Very long delays, however, probably discourage those most qualified and motivated.

**Assignment to Basic Program Activities**

Clearly, one measure of individual treatment must be the basic program activity or activities to which each individual is assigned or referred. These are normally available through MIS files, although comparability of data elements may be an issue where multiple service providers are included in the analysis. Information on multiple activities and sequencing will often require an additional data collection effort beyond the standard MIS. It is especially important to distinguish between (1) multiple-sequenced activities planned in advance, such as an orientation workshop followed by classroom training, followed by OJT in the same skill area; and (2) "second chance" activities assigned to individuals who failed to utilize their first service successfully and are more likely to fail the second time also.

**Treatment Intensity and Completion**

In addition to the type of program activity, the length and intensity of the activity should be measured, along with whether participants complete their programs. The most common measure here, length of time enrolled in JTPA, is easy to obtain from MIS files, but confounds several incompatible measures: length of planned treatment, completion of treatment vs. dropout, addition of "second chance" treatment, treatment vs. dropout, addition of "second chance" treatment, and extension of enrollment in a postprogram administrative hold while employment is
sought. For some of these factors, shorter enrollment indicates probable employment failure; for others, longer enrollment indicates probable failure. They must, therefore, be measured separately to produce unambiguous results.

Unfortunately, no precise measures of training intensity per-time-period exist. However, partial indicators of intensity can be measured once the intended nature of the intervention is identified. These might include whether the intervention results in a credential, in fields where one exists; how many hours of formal training and of "hands-on" training are provided; whether formal training is "mainstream" (satisfying state certification, taken by nondisadvantaged individuals also); and if the intervention involves employment, how complex the job is and how much new material it presents for participants to learn.

Perhaps the most important form of treatment intensity to measure is the presence or absence of multiple program activities assigned in sequence to address multiple barriers to employment. If process analysis indicates that serving the hard-to-serve is a priority for the organization, then this approach is especially important to analyze.

**Characteristics of Trainers**

Although factors such as trainers' methods or organizational arrangements can seldom be changed by JTPA, knowledge of which types of trainers most effectively produce desired outcomes can improve quality control and referral decisions. In addition, information on effective training approaches may be of interest to schools and especially to employers, since relatively little is known about how to train effectively on the job.

Some measures describing trainers can be gained from participant follow-up surveys (Simpson 1982). However, the most reliable measurement sources are trainers themselves, i.e., schools or employers. In classroom training, the easiest measures are *typologies of trainers*, e.g., trainers who enroll primarily JTPA participants vs. mainstream trainers, or trainers who are public, proprietary, or community-based. In addition, trainers vary in size, mix between experiential ("hands-on") and formal learning, inclusion of internships, and a great range of other characteris-
tics. It is advisable to construct such measures in collaboration with local vocational educators, who are most aware of variations in available training variants.

In the cases of employer-based treatment, i.e., on-the-job training and, to a smaller extent, work experience and youth tryout, a wide range of measures is available. Relatively little research has been done, however, on such measures. Simpson's (1984a) Washington State OJT research identified characteristics of the trainer and OJT positions, including employer's growth rate, typical non-JTPA turnover rate, the industrial sector, including whether public or private, the quality and complexity of the job, the use of relatively formal training methods, and range of transferability of skills gained from training, i.e., do they apply to a wide range of jobs or are they "firm-specific"?

Expenditures per Individual Participant

The primary marginal cost for each JTPA participant, is the direct cost of training. Although other costs are typically impossible to consider during a gross impact analysis, marginal training costs for each participant are usually available through contracts with trainers or employers. (See Zornitsky et al. 1985.) This means that the major program costs attached to each specific participant could be analyzed if these records can be integrated with the basic data set being used. However, such analysis is not to be confused with benefit-cost analysis. (See Johnson and Stromsdorfer, chapter 2 in this volume.)

Another cost issue that also represents an agency policy issue is ancillary support services. Except for the issue of stipends offered during classroom training, there appears to be no cogent reason to detail specific support services. However, the total amount expended per person could be recorded at little cost. Such services might affect program completion, although no postprogram impact would be expected beyond that caused by completion.

Needs-based payments are complex to analyze. Income from any source has the potential to affect life stability, personal stress, and other factors, which can in turn influence postprogram labor market experiences. Therefore, a precise analysis of the impact of needs-based sti-
pends per se requires measurement of total income during training as well as income from stipends.

Program Exit and Job Search

Agency implementation variables, discussed earlier, measure the availability of various supports at termination. Individual-level measures indicate who makes use of which support services and provide the basis for aggregate agency-level variables. Possible measures include a number of job search services: enrollment in job club or less extensive job search courses, receipt of specific job referrals from trainers or JTPA staff, and receipt of less formal job search assistance from JTPA staff. It is also useful to measure when the services occurred. If a job search workshop occurred during or prior to training, all participants have had benefit of it by the time they have to look for work. If the workshop occurs after the end of training, the most successful participants will not enroll because they will have found jobs already. This creates a “compensatory effect,” which is extremely difficult to analyze validly.

In addition, participants’ job search behaviors may be important to measure. These include the importance participants place on finding or retaining work, using the skills learned during the JTPA treatment, and the extent to which a participant is “place bound” (unable to relocate). The expressed importance of working in the training area can be used as a control variable, but can also be analyzed as an intermediate program outcome. Those intending to use such analysis may wish to measure the same variable at enrollment in order to allow an estimate of change during training.

Measurement of Control Variables

As discussed earlier, the approach differential impact analysis takes to prevent bias involves measuring “control variables” for inclusion in multivariate statistical analysis. Many of the most important protections against bias, such as measuring selection criteria, intake procedures, and exit practices, are also of interest as implementation and treatment variants. Analysis of such measures serves the dual purposes of testing their impact on program outcomes and also testing whether their exclu-
sion from multivariate equations would bias estimated impact of other program variants. Individual measures, such as preprogram barriers to employment and the route into the JTPA training activity, are also of interest for programmatic reasons, but in differential impact analysis they operate primarily as control variables.

Other control variables fall into two categories: individual background characteristics and the labor market environment. These measures are not analyzed in the hope of improving programs by changing them; most of them cannot be changed by program managers, or will not be changed, since they are part of the program mandate. They are important because they are likely to affect program outcomes and to differ across service providers, program activities, or other program variants. Therefore, unless they are measured and included in differential impact equations, the estimated impact of program variants of interest to program managers can be biased.

Several individual background characteristics affect program outcomes: inherited characteristics, such as gender and ethnicity; previously achieved characteristics, such as education level and work experience; and life cycle situation, such as marital status and number of dependents. Some mix of these measures is normally available in management information systems. Where factors known to affect labor market experiences are omitted from MIS files, or where measurement is truncated to distinguish only program eligibles from noneligibles, MIS files must be augmented. Borus (1979) provides a detailed enumeration of individual background measures found to influence employment status. Readers are referred there for information on the development of control variables measuring individual background.

One difficulty with preprogram measures is establishing their proper time frame. "Preprogram dip" in earnings and employment has been grappled with in much detail, making clear that information running back as far as three years before the program can be useful (Bloom and McLaughlin 1982; Johnson and Stromsdorfer, chapter 2 in this volume). That period may be too costly for agency data collection, but it is clear that too short a preprogram period, such as three or six months, will underestimate the long-term earning potential of many participants and will fail to distinguish those with temporary problems from others.
Measures of labor market characteristics are quite powerful in national studies. They appear to have less effect on training outcomes in a single state or locale. Even so, any study comparing service delivery in more than one geographic area must test the possibility that differences in unemployment levels, average salary levels, or demand for particular types of jobs may affect estimates of differential impact analysis. Aside from census and employment data, the availability of labor market measures depends primarily on the role each state has played in developing reliable occupational outlook data. Most SDAs have compiled this information during their planning periods.

**Construction of Category Identifiers**

The simplest but most important form of data any differential impact analysis requires is a *set of code numbers that identifies individuals and their membership in various categories*. These categories include sets of service providers, specific trainers, and types of services. Each category to be identified during analysis must have a unique identification number. These are required—

1. To merge data from different sources, allowing the construction of data sets that include the full range of test and control variables, and allowing inexpensive service provider data to be integrated into individual-level analysis.

2. To organize the data set and know what original records to consult in cases where errors on the computer file must be corrected.

3. To construct membership identifiers. These variables are the vehicle required before differential impact analysis can test the impact of membership in particular organizational units, activities, or fields. (The construction of such variables is discussed below.)

The identifiers listed below should be included in any analysis which utilizes data from each source mentioned. The precise nature of each identifier depends on the common practice in the state or SDAs mounting the analysis effort.

**Participant identifiers.** Codes identifying individual participants are the basic data-file organizing unit and are also necessary in order to merge data from MIS files, follow-up interviews, and individual treat-
ment records. The best participant identifier is a social security number, which is unique and normally required if official data such as UI, welfare, or criminal justice information is to be combined with JTPA data.

**Employer identifiers.** If employer interviews are conducted, data must be collected under an identifying code, unique for each employer, which is also recorded on each participant’s file. In this way, the appropriate employer data may be added to the files of each individual. If agencies have not yet developed employer identification codes, they will also find them extremely useful for organizing employer relations and marketing information and assessing use patterns and retention track records of participating employers.

**Classroom trainer identifiers.** If special data are collected on classroom trainers, they must be catalogued under identifiers also included on participant records to allow data merging. In addition, trainers enrolling a sufficient number of participants in a sample may be tested using either/or membership variables, if each trainer has a unique identifier.

**Training field identifiers.** The field in which participants trained or gained work experience should be identified. This allows description of outcomes by field, construction of either/or membership variables where the number of cases allows, and introduction of labor market data tied to training field.

**SDA and subcontractor identifiers.** When an analysis combines SDAs, each must be uniquely identified in order to test for differences in outcomes produced by each and add labor market data to individual computer files. The same is true for subcontractor comparisons within one or more SDAs. If subcontractors are numbered within each SDA, unique identifiers can be formed by combining SDA and subcontractor identifiers.

**Time period identifiers.** The simplest reliable way to calculate time periods, such as lag between eligibility and enrollment, is to record the date of each event, including eligibility, enrollment, treatment start, planned treatment end, actual treatment end, termination, and follow-up. If dates are expressed in compatible units, time periods can then be calculated by subtracting one from the other.
Measures That Place Special Demands on Sampling

The measures of program variants just reviewed may be placed into four categories, in order to focus on two types that make special demands on the size or structure of samples. These four approaches are shown in exhibit 3.10. The term “variable description” refers to measures that may take a range of values, for example, percentages, amounts, or degrees of some quality. These are the most common types of measures, explaining why measures are often referred to as variables. They are distinguished from “membership identifiers,” which always take only two values. The individual (1) does or (2) does not belong to that category.

Standard sampling considerations are structured for variables measured at the individual level. The two-program implementation measures in exhibit 3.10 place special demands on the sample. The demands differ depending on whether the measure is a membership identifier, indicating whether or not a participant was served via some specific treatment context, or a variable description of some particular aspect of program implementation. Membership identifiers, which indicate whether or not individuals received services via a particular organizational or treatment activity provide an effective way to locate impacts on postprogram outcomes, but not to explain why they are located where they are. Variable descriptions of program characteristics, on the other hand, measure specific qualities that vary across all service providers or program activities, rather than separating each as a whole from the others. This approach does not pinpoint concrete contexts where differ-

---

**Exhibit 3.10**
Four Approaches to Measuring Program Variants, With Examples

<table>
<thead>
<tr>
<th>Program Implementation (Measured via a Survey of Service Providers)</th>
<th>Measure of Individual Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifier of membership in a specific context</td>
<td>Enrolled through N W corner SDA, subcontractor No. 3 versus all others</td>
</tr>
<tr>
<td>Variable description of program characteristics</td>
<td>Percent of services performed in-house</td>
</tr>
</tbody>
</table>
ences occur. However, it helps explain why they occur, a quality which makes them especially helpful for program development. The particular demands placed on sampling differ between the analysis of membership identifiers and the analysis of variable program descriptions.

**Membership Identifiers**

Measures that identify membership in particular treatment contexts place little restriction on the number of service providers or other treatment contexts required. If two contexts are identifiable, they may be compared by entering the membership identifier into an equation that also includes the appropriate control variables. However, this simplicity is gained at some cost. First, *explanatory power is no greater with 20 contexts than with two.* Each context is compared individually with all others. Second, *reliability of such comparisons depends on the number of participants enrolled in each membership group.* Thus, no matter how large the total sample, estimates for membership in a category containing very few participants cannot be reliable. This means that analysis of this type may require large, disproportionately stratified samples.

Membership in highly specific contexts, such as particular schools or employers, is usually immune to analysis because so few individuals belong to each context. Membership in larger units, such as SDAs, is entirely possible to assess given that a large enough sample is drawn from each SDA. This limitation on the analysis of specific treatment contexts is of greatest interest for states and large SDAs, which may wish to assess relative performance among different units within the system.

**Variable Descriptions of Program Implementation**

The sampling demands made by variable descriptions of implementation are directly opposite to those made by membership identifiers; they require multiple treatment contexts, but not larger samples within each. Since participants in each context receive a specific value on some measurement scale, the number of participants in each context matters little. Thus, sample stratification is unnecessary. That advantage, however, is purchased at the cost of requiring *multiple treatment contexts* to be included in the sample.
Imagine that two service providers have been measured on two variables—intensity of intake procedures and the degree of job search assistance provided. If we find that the two providers differ in outcome level, how can we decide which of these variables accounts for the difference? For that matter, how can we claim that either of these, rather than some other variable, explains the difference? To assess whether intake or placement had the impact, we need to compare situations characterized by thorough intake but little job search assistance, and vice versa. With only two organizations, however, that is not possible. These two agency characteristics, as well as any others one can imagine, are by definition perfectly correlated and cannot be disentangled. (In statistical terms, only one degree of freedom is available.)

This same problem faces research comparing more than two contexts, where the variable in question happens to differentiate only one from all others. An analysis reported by Franklin and Ripley (1984) illustrates. They report that program performance was lower in CETA prime sponsors characterized by "crisis management" style. While this finding appears reasonable, only one prime sponsor in their sample was so characterized. Their conclusion, therefore, was based on a comparison between one prime sponsor and 14 others. Consequently, any number of other qualities of that one prime sponsor could have produced the differences they observed.

In the case where three service providers are included in a sample, it is very likely that the problems discussed above will remain. However, there is now a possibility that, in unusual circumstances, one variable characteristic of service providers would have such a strong and consistent impact that a statistically reliable effect would emerge. The principle of parsimony—using the simplest explanation consistent with the facts—becomes the guide to interpretation here. If the differences among outcomes in the three contexts closely fit a single linear treatment measure, but no others, then it is parsimonious to explain findings with that one factor. If, however, they vary far from a linear fit, or if more than one variable fits equally well, the less tidy, but more accurate interpretation must be used; namely, that each unit differs from the other for reasons we cannot demonstrate.
If we introduced a second treatment characteristic variable into the analysis based on three contexts, we would automatically revert to the case in which it is impossible to distinguish among competing explanations. In statistical terms, the number of variables that may be uniquely estimated may not be greater than the degrees of freedom, which equal the number of cases minus one. Since these variables are measured only at the organizational level, the number of cases we are speaking about is the number of service-providing organizations in the analysis.

Extending this line of thought, it is apparent that analysis including many SDAs or analysis of large SDAs including many service providers can be especially valuable for local program development. The larger the number of different agencies in the analysis, the more feasible the tests of agency implementation variables. More variables can be handled simultaneously, and each is tested more reliably and less ambiguously. That is, other things equal, the more separate service providers included in a sample, the lower the covariance among implementation variants is likely to be, strengthening the ability of multivariate analysis to estimate the unique effects of each.

This general rule, that the larger the number of contexts, the firmer the analysis of variable program characteristics, leads to a practical question: what is the minimum number of service providers required for a reasonable differential impact analysis of agency-level implementation measures? The answer is twofold. First, the bad news. The answer depends on many factors: variance in each independent variable, variance in the outcome variable, covariance among independent variables, and covariance between independent variables and the outcome variable. Therefore, no precise minimum can be set forth. One might reasonably say that there is little point in pursuing analysis of variable program implementation measures with fewer than six or seven service providers. In many cases this would be too few, while in others, it would be sufficient.

Second, the good news. There is an analysis procedure that can in most cases protect against incorrectly attributing too much importance to variable descriptions of program characteristics. This procedure involves jointly testing the variable program characteristic measures along with membership identifiers indicating enrollment in each particular
service provider in the sample. After identifying variable program characteristics that appear statistically reliable, the analyst then adds to the equation the set of membership identifiers indicating enrollment through each specific service provider.  

If the variable program characteristics retain their statistically reliable effects, then our confidence in the initial findings remains high. If their effect in the equation is eliminated by the addition of the membership identifiers, then we must conclude either that some service providers differ from each other, but we do not know why, or that the initial test procedure was inappropriate. With small numbers of units, the latter is always a strong likelihood.

Analysis Procedures for Descriptive Outcome and Differential Impact Evaluations

This chapter makes no attempt to provide instruction in the use of statistics. However, a brief overview of analysis strategies for descriptive gross outcomes analysis and differential impact analysis may be useful.

Descriptive analysis involves quite basic statistical tools. The value of descriptive analysis rests more on the thought that goes into the questions the analyst asks than on statistical sophistication. Descriptive analysis begins with univariate (one-variable) averages or percentage distributions. Beyond that, bivariate (two-variable) associations can be calculated, as long as the analyst bears in mind that descriptive associations can be produced by many factors other than the two being analyzed. Exhibit 3.11 describes conditions under which different bivariate statistics are most appropriate.

Differential impact analysis can be performed satisfactorily with standard multiple regression techniques, except for one particular situation, which is discussed. The strategy of multivariate analysis is straightforward. One outcome is analyzed, with multiple potential influences tested simultaneously to estimate the unique impact of each on the outcome. In this instance, the goal is to ascertain whether and how much policy variables of interest affect the outcome after taking into account the possible effects of other factors such as selection. However, the statistical techniques required to implement that strategy require special-
ized training. Parts of the discussion that follows assume prior background in multivariate analysis.

**Analyzing Various Types of Measures**

A wide range of statistics is available in various software packages. However, nearly all statistical tests required for descriptive gross impact analysis or differential impact analysis can be performed with four basic tools: chi square, analysis of variance (ANOVA), Pearson correlation, and ordinary least squares (OLS) regression. Why these are typically adequate is laid out in a highly readable form in Bjornstadt and Knoke (1982). Which of these is used depends on the nature of the analysis and the way in which the variable was measured—commonly referred to as the level of measurement. Exhibit 3.11 suggests appropriate statistics for different levels of measurement and for different analysis goals.

The critical distinction regarding level of measurement is between ordered and nominal variables. Ordered variables are those for which values assigned to each category of the variable form a logically defensible sequence from smaller to larger, lower to higher, etc. Ordered variables include age, level of satisfaction, costs, ratings on various descriptive scales, and the like.

Variables that cannot be ordered are termed nominal variables. The categories of variables like marital status or SDA identification codes cannot be placed in a meaningful hierarchy or sequence. The results of tests that require ordered variables would be meaningless if applied to nominal variables such as these.

Dichotomous variables, those taking only two values, such as “yes” and “no,” occupy a special status in that they are by definition ordered, even when they appear logically nonorderable. Any variable that includes only two values can be expressed as a yes/no question. In the case of one SDA vs. others, for example, the variable becomes “Did this participant enroll through SDA #1?” The responses “yes” and “no” are interpretable as ordered, with yes greater than no. It is this quality of dichotomies that makes membership identifiers especially powerful in differential impact analysis.

Statistical assumptions vary somewhat for dependent (outcome) vari-
### Exhibit 3.11
Suggested Statistics for Different Levels of Measurement

<table>
<thead>
<tr>
<th>Type of Analysis</th>
<th>Type of Variable</th>
<th>Suggested Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Independent (e.g., a program variant)</td>
<td></td>
</tr>
<tr>
<td>Bivariate</td>
<td>Ordered</td>
<td>Ordered</td>
</tr>
<tr>
<td>Bivariate</td>
<td>Ordered</td>
<td>Dichotomous</td>
</tr>
<tr>
<td>Bivariate</td>
<td>Dichotomous or nominal</td>
<td>Ordered</td>
</tr>
<tr>
<td>Bivariate</td>
<td>Dichotomous or nominal</td>
<td>Dichotomous or nominal</td>
</tr>
<tr>
<td>Multivariate</td>
<td>Ordered or dichotomous</td>
<td>Ordered</td>
</tr>
<tr>
<td>Multivariate</td>
<td>Ordered or dichotomous</td>
<td>Dichotomous</td>
</tr>
<tr>
<td>Multivariate</td>
<td>Any</td>
<td>Nominal</td>
</tr>
<tr>
<td>Multivariate</td>
<td>Nominal</td>
<td>Ordered or dichotomous</td>
</tr>
</tbody>
</table>

¹ If ordered variables contain few (3 – 6) categories, it may also be advisable to observe relationships in tabular form. However, the chi square statistic would typically underestimate the likelihood of a reliable relationship because it ignores information on order.

² In this case it is convenient to treat the independent variable as the dependent, and vice versa, so that ANOVA may be used.

³ In the dichotomous case, the t-test is equivalent to the F test used in ANOVA.

⁴ If available to the analyst, recent developments by Goodman (1972) make limited multivariate analysis of nominal variables possible (See also Davis 1974.) Goodman’s program is named ECTA (Everyman’s Contingency Table Analysis). SPSSx has also installed a version.

ables vs. independent (predictor, explanatory) variables. Therefore, the choice of statistical tools depends on the level of measurement for each. Exhibit 3.11 reflects this requirement. Analysis goals are separated into bivariate (two-variable) and multivariate (one dependent variable, more than one independent variable) cases, with measurement indicated for both independent and dependent variables.
Multivariate Analysis with a Dichotomous Dependent Variable

Every statistic is developed on the basis of mathematical assumptions. In the case of ordinary least squares regression, many of the original restrictive assumptions have proven unnecessary. That is, the statistic is highly robust; data can be structured in ways not fully satisfying statistical assumptions, yet the statistic produces accurate and efficient estimates. Even so, in the case where the dependent variable is dichotomous and is highly skewed (unevenly distributed), assumptions are violated severely and error can result.

Happily, recent work with statistics based on log-linear transformations of dichotomous dependent variables—and using "maximum likelihood chi square goodness of fit" tests—avoid the problems faced by ordinary regression. This means that appropriate conservative multivariate methods to analyze dichotomies, such as whether or not participants are employed at follow-up, do exist. Regression tests have been compared with these more conservative methods, with the result that we can now be quite certain when we are required to use the more conservative, but also less convenient, methods and when the simpler regression analysis is appropriate. (See Knoke 1975; Goodman 1976; Gillespi 1977.) The following guidelines summarize this knowledge:

1. If a dichotomous dependent variable is split relatively evenly (between 75 percent/25 percent and 25 percent/75 percent) OLS regression may be used.
2. If OLS regression cannot be used and all or many independent variables are ordered, Logit or Probit transformations of the dependent variable are advisable.
3. If OLS regression cannot be used and many independent variables are dichotomous or nominal, Goodman's Multiway Contingency Table Analysis may be used.

Constructing and Testing
Membership Identifiers ("Dummy Variables")

Some extremely valuable factors to include as independent variables in multiple regression analysis are measured as nominal variables, which are not admissible in a regression equation. Nominal variables may be
analyzed, however, after they are transformed into dichotomies that are tested in place of the original variable. These dichotomies, known as "dummy variables" in formal statistical analysis, are called membership identifiers throughout this chapter to indicate their nature, i.e., they measure whether or not a participant belongs to a particular category. For example, MIS systems may include a measure of ethnicity, including values for each of five or more major groups. If membership variables are constructed for each of these groups (e.g., a variable named "othwhite," and scored 1 for all originally coded "other white" and 0 for all others), these new dichotomies may be tested as independent variables in regression analysis.

Regression slopes that result from tests of dummy variables, if statistically reliable, indicate that members of the named group (e.g., those enrolled through unit A, "other white" participants, or OJT participants) are the estimated average amount higher or lower on the outcome in question than all members of other groups. All but one of the dummy variables created from an original nominal variable may be tested in a single equation.

**Constructing and Testing Interaction Terms**

One useful type of question for JTPA program managers may be addressed by using interaction terms. This question is Do particular groups of participants, more than others, experience greater success from some program variants than from other variants? For example, is classroom training (CT) more successful than other treatments in erasing the deficit produced by previous low educational attainment? In an interaction, two variables combine to produce a joint effect different from that which both acting independently would produce. In the example, dropping out of high school reduces postprogram outcomes and CT may, in itself, produce higher or lower than average outcomes. In addition, the interaction hypothesis suggests that the impact of educational attainment on postprogram outcomes is stronger when the treatment is not classroom training than when it is. To test such an hypothesis, one must construct an interaction term that would be scored 1 for dropouts who enroll in CT and 0 for all others. This interaction term,
the product of the other two, identifies those individuals who were in a position to have some portion of their educational deficit eliminated by participation in CT.

Normally, the interaction term, appropriate control variables, and the original variables from which the interaction was constructed must all be included in one multiple regression equation (Blalock 1965). Since the interaction term will include portions of both original variables, an effect of either or both original variables would erroneously be carried by the interaction term alone. Only when the original terms are both included in the test can we be certain an effect of the interaction term is not spurious. If such a test produced, for example, a slope of -0.20 for dropout status and +0.10 for the interaction term, then the proper interpretation would be (a) that being a high school dropout, in itself, reduces postprogram employment 20 percent, and (b) that CT erases half of the education effect among dropouts, so that dropouts who enrolled in CT experience only a 10 percent lower placement rate (-0.20 +0.10).

Reporting Standardized or Unstandardized Regression Coefficients

Standardized regression coefficients, termed “Betas,” are often reported because they indicate the relative power of each variable in one equation to account for variation in the dependent variable. Betas have a commonsense meaning similar to that of a correlation: a Beta of 0.5 always indicates a “stronger” effect than a Beta of 0.4. Unstandardized coefficients (regression slopes) are expressed in terms of the metric of the independent and dependent variables, and are much more precise but often less intuitively satisfying. If, for example, education is scored using a four-point scale, a slope of 0.10 indicates that each step of that education scale raises the outcome variable by 0.10. The lowest step compared with the highest, three intervals above, has an estimated 0.30 higher level. If the dependent variable is employment status, 0.30 translates to 30 percent. If it is hourly wage, 0.30 translates to 30 cents.

These considerations make slopes somewhat more complex than Betas to communicate when findings are reported. When results of research are being applied to program development efforts, however, unstandardized slopes are the preferable estimate because they give a
direct estimate of the amount of change in an outcome that is produced by a given change in the input. Is it more helpful for me to know that education, which I cannot control, is more powerful than program activity? Or am I better served by estimating that after the effects of education are accounted for, my OJT program produces postprogram wages an estimated 47 cents lower (or higher) than my CT program? Clearly the latter, unstandardized report is preferable.

In addition, unlike standardized Betas, unstandardized slopes are not influenced by the variance of a particular variable within a sample. This can be important when relatively small subgroups are being analyzed. If only 10 percent of my sample enrolled in work experience, even if WEX participants are employed only half as often as others, the variable “enrolled in WEX?” can explain only a limited portion of the variation in employment experienced by the entire sample. However, the unstandardized slope indicates how much less often this minority is employed than are other participants. The slope remains the same whether 10 percent or 50 percent of participants are enrolled in work experience.

In general, unstandardized regression slopes are both more useful and easier to report when (1) the dependent variable is naturally interpretable, as in the case of income or a dichotomy that translates to percentages, and (2) the independent variable is a dichotomy, allowing statements like “participants in category A are X percent higher than those in category B.” In other cases, the analyst must choose between reporting ease and managerial usefulness. For a full analysis, both forms augment each other. For example, a report might indicate that a particular program service has a negligible impact on variation in outcomes for an entire SDA, but go on to show a large impact on a few clients.

**Estimating Change**

For descriptive analysis, change may be indicated by subtracting the preprogram value from the postprogram value of an identically measured variable. This procedure is simple and the results are often taken at face value. However, the descriptive report of change is especially problematic because change is heavily dependent on the original base figure. If, for example, a sample includes many students or displaced homemakers
with zero earnings during the preprogram year, then regardless of program impact, that group is likely to generate higher change in earnings than a group including primarily high previous earners. At the other end of the scale, displaced-worker programs will typically show negative change figures—a reduction in earnings—not because they are less effective than other programs, but because preprogram earnings were especially high. This problem is especially severe if preprogram values vary more widely than postprogram values, as is more often the case in highly successful programs than in less successful programs. For this reason, presentation of descriptive change results can be seriously misleading.

In multivariate differential impact analysis, the goal is to estimate unique causal effects of each factor tested. Since the preprogram level of an outcome variable clearly affects change in that variable, any analysis of change must include that level as a predictor in the regression equation. This necessity leads to an approach that considerably improves the ease and usefulness of change analysis. Predicting change with preprogram level produces awkward results, i.e., higher preprogram levels of any variable produce lower rates of change. However, it is possible (and preferable) to replace change with postprogram level as the outcome being predicted. Since change is calculated as postprogram minus preprogram level, any two of the following variables is sufficient to produce the third via simple mathematical operations: (1) the preprogram level, (2) the postprogram level, and (3) change. Therefore, the equation is satisfied regardless of which two are used.

It is preferable to use the two that produce the most sensible results: the postprogram outcome as dependent variable and the preprogram measure of the same outcome as a control variable. In such a case, the effect of preprogram on postprogram level (the autoregression term) indicates stability over time, i.e., the tendency for those most employable before the program to be most employable after the program also. Other variables in the equation that show a reliable impact on the postprogram outcome may be correctly interpreted as indicating factors that increase or decrease (change) the outcome in question from the preprogram to the postprogram period.
These considerations open another possibility. It is a short step to conclude that the multivariate analysis of change is hurt very little if the preprogram measure differs slightly from the parallel postprogram measure. The analysis still indicates which factors increase or decrease the outcome controlling for preprogram level of approximately the same factor. Identical pre- and postprogram measures are desirable. However, the analysis of change remains possible with less-than-exactly parallel pre- and postprogram measures.

**Analyzing a Large Number of Potential Influences on Outcomes**

Multivariate analysis is straightforward in studies where only a few theoretically derived variables are tested. All are entered into the equation and the results reported. In cases where many measures are to be tested as independent variables, however, it is no longer possible or advisable to include all in a single test. Such attempts can make undue demands on the sample size, a problem that becomes especially serious if the sample size is reduced by the accumulation of missing cases from each of the many variables involved. This produces the problem of how to move through multiple tests efficiently without distorting or overlooking effects. The following suggestions may assist in that endeavor. Performing such analysis, of course, requires prior statistical background. This “cookbook” summary is not meant to imply otherwise. It only suggests steps to make analysis relatively efficient.

**Step 1.** Insure that variables are in the proper form for multivariate analysis and that variation is sufficient to make analysis meaningful. For data management purposes it is advisable to construct a “codebook” listing all variables and showing for each the level of measurement, number of useable cases, and an indicator of variation.

**Step 2.** Select the appropriate dependent variable for each analysis.

**Step 3.** Separate variables according to their importance to the analysis. Those that are most important, because they are known to affect the outcome and must be included to prevent bias or because they hold special program development interest, should be given priority during analysis.

**Step 4.** Separate variables according to missing cases. In particular,
questions asked only of subsets of participants, such as job qualities or reasons for unemployment, should be analyzed separately from other questions applying to all. The safest form of multivariate analysis is based only on cases for which full information on all variables is present. Under that approach, any case with a single missing value is eliminated from the entire analysis.

Step 5. Compute correlations between all independent variables and the dependent variable being analyzed. Correlations are the basis for the calculation of multiple regression coefficients, making them the appropriate bivariate test building toward regression analysis.

Step 6. Identify those variables that are appropriate in terms of missing values and have high priority as control variables—those required to protect against biased estimates. Observe their correlations with the outcome(s) in question. Select from this set those variables exhibiting a reliable association with the outcome being analyzed.

Step 7. Enter the variables selected at the conclusion of step 6 into a multiple regression equation, to identify the subset of these variables that have reliable multivariate effects on the outcome. For simplicity, analysts may use a stepwise procedure, which automatically selects reliable effects. This produces a minimal set of control variables which must be included in subsequent runs. Other variables from these tests may be set aside for the moment with the knowledge that were they included in the regression equation, their effect would be too small to alter findings noticeably.

Step 8. Identify the most important test variables, i.e., program variants of special policy interest. Observe their correlations with the outcome in question, selecting those showing reliable association. Enter these singly, or in appropriate sets, into equations that include the minimal set of control variables identified in step 7.

Step 9. In addition to variables tested in step 8, analysts may wish to explore other program variants, hoping to discover useful unexpected relationships. Group measures according to policy area, such as intake, quality control, or trainer characteristics. Correlate these with the outcome being analyzed and enter those which are reliably greater than zero into an equation including the minimal set of control variables identified in step 7.
Step 10. The procedures outlined above for reducing the set of reliable effects ignores the possibility of suppression, a situation in which two variables are correlated with each other but have opposite effects on the dependent variables and, therefore, tend to cancel each other out in bivariate tests. These effects become visible only when both independent variables are tested jointly. They are, therefore, overlooked when only reliable bivariate correlations are forwarded for test in regression equations, as in steps 6-9.

Short of a full exposition of this issue, one step may be suggested to guard against most errors of this type. Suppression of the type the analyst most wishes to uncover occurs only when some variable is correlated with one of the variables identified during steps 8 and 9 as reliable predictors. Correlations should, therefore, be calculated between each of these reliable effects and other independent variables. Where reliable correlations are found, the variable in question may be added to the reduced set of reliable effects located after step 8 or 9. Relatively few changes will be produced by such a procedure, but it does guard against the most damaging errors from undetected relationships. These tests may be facilitated by using a backwards stepwise elimination of unreliable effects, where statistical packages include this option.

Step 11. Membership identifier variables, such as service provider, industrial sector, and similar others, should be examined if they have not already been included as program variants. These may be added to the reduced sets of reliable effects identified in steps 8, 9, or 10. Findings may prove useful for future contracting or marketing. Also, such tests protect against spurious findings of program effects.

Step 12. Finally, having identified reduced sets of the most powerful and unique effects on each outcome being analyzed, the analyst will be well-advised to return to the data set in order to examine what measures are associated with these key effects. Such analyses may be conducted formally, using these key effects as dependent variables in their own right, or may be undertaken as less formalized examinations of patterned associations. Such further analysis can corroborate or challenge initial interpretations, or can help the analyst develop interpretations of initial findings by detailing the apparent nature of the variables found to have greatest impact on the outcome.
Combining Data from Different Sources

One complexity of effective differential impact analysis is the need to combine data from several different sources into a tailor-made data set. The combination of data also provides two of the method's strengths: the ability to protect against selection bias from several angles, and the ability to measure different types of variables in the most reliable and efficient manner.

Aside from the availability of appropriate computer facilities, the key to combining (merging) data is to include the correct identifiers in each data set to be merged. It is advisable to produce a master identifier cover sheet to become part of each participant data file. This sheet should include all the identifying information required to merge data: participant social security number, SDA identifier, agency (subcontractor) identifier, employer identifier, etc.

For differential impact analysis, all data should be merged into individual participant records, since the participant is the unit of analysis. All identifiers must appear in the participant's original data file. Each of the other original files must contain only the particular identifier required to correctly merge into the participant file. For example, implementation program variants are measured at the agency level. Each participant who enrolled through the agency with the ID code "10" will receive values on all implementation variables which were provided by that agency. The agency identifier will appear on those participants' master identifier sheets and also on the appropriate agency implementation data reports, allowing the match of identifiers, followed by the combination of data.

Once data sets are merged, statistical tests will be calculated on the basis of the number of participants (or employers) in the data set, not on the basis of the number of service providers or geographical regions that may have supplied particular data elements. The analyst must, therefore, remain aware of limitations surrounding the number of separate treatment contexts required for reliable differential impact tests (discussed earlier).
Research Design for an Employer Follow-Up

With the advent of JTPA and the expansion of private sector involvement, interest in measuring employer benefits has risen. The popularity of this issue among service providers is no doubt connected to a concern for marketing JTPA services and products to employers. The perceptions of OJT or WEX employers are useful to indicate which program approaches are relatively effective, which are distasteful to employers, and what steps might encourage or discourage future participation by employers (e.g., Wentling and Lawson 1975; Minnesota Office of Statewide CETA Coordination 1979; Simpson 1984b). Employers may be interviewed primarily as a marketing tool. In addition, employer interviews can be valuable program evaluation tools. They may be used to assess the effectiveness of participant services or to assess employer benefits, both central to JTPA program development. How these goals translate into a research design depends on the relation of the employers being studied to JTPA, in particular, whether they are participating employers or termination employers.

Termination employers are those who employed participants at their termination from any program activity. They are consumers of JTPA’s “products.” Some hire former JTPA participants without knowing that the training their new employees received was supported by JTPA. Others have participated in providing training or experience to the participants they subsequently hire at termination.

Participating employers are those who participated in the delivery of services, through on-the-job training, work experience, or tryout, regardless of the termination status of the participants involved. Many become termination employers also. However, many employers participate in contracts that end prematurely, or complete a contract but choose not to hire the JTPA participant following his or her participation in JTPA.

For termination employers, surveys may ask direct marketing questions, such as how the employer came to hire a JTPA product, or may address indirect marketing goals, such as measuring employer satisfaction with the former participant hired, with the goal of demonstrating program success to future employers. For participating employers, marketing questions may be expanded to include willingness to continue
or expand future participation. In this case, the goals of marketing the program and assessing employer costs or benefits overlap. The major marketing tool for engaging new employers in service delivery is evidence that past participating employers feel that the benefits of participating have outweighed the costs.

For termination employers, the goals of assessing participant services and employer benefits are nearly identical. Employer costs or benefits occur entirely because participants are or are not well-prepared for their jobs. In the case of participating employers, reports of costs and benefits can include not only ratings of the participant, but also evaluations of the JTPA programs and personnel and direct perceptions of participation as beneficial or costly.

Employer surveys are unique in that many employer reports may be taken at face value and are, therefore, especially useful for descriptive analysis. The employer’s role as consumer of JTPA products (participants), makes employer ratings of participants valuable to JTPA program operators regardless of the factors influencing them. Because JTPA agencies wish to have the most effective participating employers repeat their involvement, employer satisfaction with agency policies or personnel is critical. Similarly, employers’ perceptions that they have benefited from participating in delivering JTPA services are meaningful descriptive estimates of employer costs and benefits.

In addition, information gathered from employers can be valuable to the analysis of participant outcomes. Employers are in a unique position to report subtle job quality outcomes, certain forms of selection, and for participating employers, individual-level measures of training provided or other qualities of the program intervention.

**Identifying Employer Population to Be Analyzed and Designing the Sample**

The most basic of all employer design questions is whether the population being studied includes all termination employers, all participating employers, or both. In addition, managers who wish an in-depth analysis of one specific program activity may prefer an even more specific definition, such as all OJT employers. Aside from modest differ-
ences in cost, these decisions should be made on the basis of policy objectives. To which programs do managers wish to apply the results? Are specific services earmarked for further development? Is descriptive material on the range of all employers’ experiences needed? The research design should be shaped by these decisions; it should not drive them.

Any analysis of participating employers must include all participating employers within the program services to be analyzed. In the event that one participant is placed with more than one employer before program termination, both employers must be included. To identify the sample of participating employers only from the population of those who were the “final” employers, is to bias the employer sample by eliminating a group of placements that worked out especially poorly—those that ended prematurely and were followed by transfer to further treatment.

Sample Size

Only one issue differentiates participant and employer sampling with regard to sample size: the completion rate for employer surveys will probably reach 80 percent or more. Therefore, the initial sample of employers required to produce a target sample of completed interviews is smaller than that required with participant samples. For example, if we decide to aim for 400 completed interviews and expect an 80 percent completion rate, an initial sample of 500 will suffice. For a participant survey with 70 percent completion rate, the figure would be 571.

Integrating Employer and Participant Samples

Combining employer and participant data is recommended for any but the most basic marketing study or descriptive analysis of employer benefits. If both employer and participant follow-up analyses are conducted, samples should overlap as much as possible. The validity of each in no way depends on the degree of overlap between the participant and employer samples; it depends on the representativeness of each sample. A sample of participants selected at random will produce some proportion with employers—those employed at termination, or those in employer-based programs, depending on whether the employer survey is of termination or participating employers. This sample selection also
produces a random sample of the program's employers, and insures a substantial number of cases in which both participant and employer data are available for joint analysis.

**Follow-Up Period**

Employer surveys should allow a lag time after hire, long enough for the employer to observe the former participant's work and decide whether to retain the individual, but short enough to allow employers to retain clear impressions of participants who remained on the job only briefly. Lag time should also be short enough for impressions of working with JTPA to remain salient. Since most positions gained by JTPA graduates have relatively short probationary periods, a three-month follow-up survey should be adequate for an employer survey. This also gives participating employers who hire participants at termination enough time to observe their new workers under full-wage conditions. In cases where OJT positions involve a posttermination performance payment, follow-up should occur at least one month after the final performance payment to avoid distortion from expected payments.

**Data Collection Methods**

Most employers rely heavily on telephone communications and respond well to telephone interviews, especially if they are scheduled beforehand. Agency personnel who work with employers may resist interviewing, feeling that the intrusion on employers' time jeopardizes good will. Experience with CETA and JTPA surveys shows, however, that brief interviews are usually accepted and the majority of employers are pleased that JTPA staff care enough about the quality of their program to check with those who consume their products.

**In-House vs. Third-Party Data Collection**

Because JTPA staff work closely with many employers, there are program development advantages in having staff conduct the interviews. These interviews are efficient because they occur along with other employer contacts. They also allow staff to enhance their program development and employer quality review by integrating them with
employer data collection. Staff interviews may introduce response bias problems, however, because participating employers may wish to participate again and may be less than candid about their costs and benefits from participation.

Viewed from the standpoint of measurement validity, research efforts must be neutral. If the results of an employer analysis are to be disseminated publicly, both the employers responding to the survey and the research consumers must be assured of the neutrality of the measurement and analysis, and of the confidentiality of individual responses. Similarly, if employers themselves are being assessed, a third-party research team should collect data, with guarantees of confidentiality. If employer surveys are conducted in-house, efforts can be made to ensure the perception of neutrality (see Dillman 1978; Bradburn 1983); however, these measures cannot successfully emulate third-party neutrality.

**Estimating Employer Costs and Benefits Using the Gross Impact Approach**

Job training programs have an impact on employers as well as on participants. Employers may be viewed as direct beneficiaries of the job training system and in some cases, as incurring costs of providing services to that system. In fact, it is often difficult to separate benefits to employers from benefits to participants and society. When a placement works well, all benefit. When an employer provides training, the participant can become more employable (either within the firm, or generally), and the employer can gain a more productive worker. Similarly, the wage subsidy employers receive is rewarding to them and also to the participants, who receive full pay for a period of partially subsidized work. For measurement purposes, employer costs and benefits may be treated as if they accrued only to employers. However, interpretations of research findings should recognize that the most effective systems can probably benefit all actors—employer and participant benefits need not be mutually exclusive.

Employer outcomes are not specified in JTPA legislation. Nor is there a long tradition of past research focusing on and defining them. Indeed, our initial directions in exploring possible measures are the result of two
major limits to measurement and research design in this case: the lack of prior development in this area of research, and the inability of gross impact research to estimate net impacts.

Employer Estimates of Their Own Costs and Benefits

Within the limits of the gross impact approach, the true net impact of JTPA on employers cannot be estimated. That would require comparisons with employers hiring non-JTPA participants. Nevertheless, we can ask employers to give us their estimates of their costs and benefits. This may be accomplished by specifying a break-even point for each measure of cost or benefit, and asking employers to report whether their experience with JTPA or with specific JTPA participants fell above or below that point. The break-even point differs by type of measure and is discussed below. The strategy in each case is to express the measure in terms simulating true or perceived net cost or benefit, by wording the measure in terms of break-even point and offering responses on either side of that point. 14

Although such an approach is far from true net impact, it may provide knowledge of employer outcomes and how they combine or offset each other. This, in turn, will build a knowledge base required before conceptualizing a net impact analysis. In particular, we can analyze the degrees of association among different measures of cost or benefit and assess the relative importance employers assign to these factors. However, the main value of this approach is that it allows approximations of employer costs or benefits useful for guiding program development.

Measurement Strategy in a New Area of Study

As a relatively new area of study, employer benefits cannot be measured definitively. It is possible to specify a range of probable benefits and costs, but too little is known about each or about their relative worth to employers to develop a precise, meaningful accounting. Some of these costs and benefits, such as the OJT wage subsidy, can be expressed in precise monetary terms. Others may be equally important to analyze, but impossible to quantify or even to conceptualize clearly. 15 They include the following:
1. The major indication that hiring a JTPA participant was rewarding to an employer is a decision to retain the participant. It might also be possible to estimate how far above or below a break-even point (a point of indifference, neither costly nor rewarding) each participant’s performance falls; however, such estimates would remain speculative.

2. Fear that JTPA participants may have serious problems not easily observed before hire is impossible to quantify in precise monetary terms, but is a very important cost for many employers (Simpson 1984b). Even when no problems arise, the perceived risk that they might occur represents a cost.

3. Provision of training is costly to participating employers. Assigning quantitative values to employer training, however, is difficult because most training is informal; more training may be planned than occurs; most training would be offered to all new employees regardless of JTPA involvement; and much of the training may be so specific to the particular employer that it binds the worker to that job rather than transferring to other employment situations, introducing a hidden benefit.

Other elements of the JTPA program are complex to conceptualize because they may act as either costs or benefits. For example, employee screening can be a service to employers, but giving partial control over screening to an agency whose goal is serving the disadvantaged may be costly. Similarly, hiring the disadvantaged is typically assumed to be one cost to participating employers. Yet, one study of CETA OJT employers found that over one-tenth of the respondents listed the knowledge that “you are helping others with need” as the major reason for participating in OJT (Simpson 1984b).

We face these measurement challenges primarily because little work has been completed in this area, and many of the most important costs and rewards to employers are inherently perceptual and, therefore, not readily susceptible to monetary quantification. The approach suggested by gross impact analysis is to develop multiple measures of potential costs and benefits to employers and investigate the extent to which each is perceived by employers to act as a cost or a benefit in their specific cases. The following are some examples:
1. Once we learn which aspects of JTPA employers estimate to be most costly and most rewarding and how important they perceive different costs and benefits to be, analysts can begin to define and prioritize employer outcomes.

2. We can analyze whether particular types of employers have different perceptions of the costs and benefits of JTPA, and whether these ideas are associated with greater or lesser program success for participants.

3. We can test ideas about the ways in which JTPA is rewarding or costly to employers. Rather than assuming that particular JTPA services, such as client screening, are costly or rewarding to employers, we can examine the extent to which the implementation of these services increases or decreases the rewards or costs perceived by employers.

4. We can analyze the association among different measures of cost and benefit. Are costs of providing services higher where benefits, such as the subsidy to participants' wages, are higher? Do employers who receive high levels of one type of benefit tend to receive less of others, or is JTPA implementation such that some agencies reward employers across-the-board more than others do?

**Outcomes for Termination Employers**

Any employer's major costs or benefits from hiring are the job performance qualities of the new employee. For termination employers who did not participate in delivering JTPA services, this is the only source of cost or benefit relevant to JTPA. The question for them is whether the new JTPA-trained employee will function in the job as well as other appropriately trained new workers. There is no reason to expect JTPA participants to be better trained than others; the goal of JTPA is to eliminate participants' previous deficits.

Each area of worker performance of importance to the employer represents one dimension, or scale, of cost or benefit. How many days' work will the new employee miss during the first month? How much employer training will be required before the worker becomes productive? How much supervision time will be saved by a "self-starting"
worker? Cost and benefit represent two ends of each scale. The break-even point lies at the point on the scale that represents the average new (non-JTPA) hire for that job in that labor market, as perceived by the employer. If the average new-hire misses four days' work per month, hiring a former JTPA participant who misses an average of two days represents a benefit to the employer.

For employee qualities that are not naturally quantified, the former JTPA participant may be compared with the average non-JTPA hire for that same job, using a rating scale such as "much better," "a little better," "about the same," "a little worse," "much worse." Qualities to measure include skill level, speed and quality of work productivity, indicators of supervisability, and indicators of adjustment to the job.

In the case of any one participant, job performance may be better or worse than average for reasons unrelated to JTPA participation or referral. However, if, over a large number of employer interviews, the average JTPA hire proves to be more satisfactory to employers than their average non-JTPA hires, we have reason to claim a role for JTPA in producing that benefit to employers.\textsuperscript{16}

A second possible benefit to termination employers is a former participant's job retention record throughout the follow-up period, and whether further retention is likely. Retention implies that the worker is productive and adjusted, and also wishes to remain employed. Unless job loss results from cutbacks forced by declining business, laying off a trained worker indicates a cost to the employer: the cost of hiring and retraining another worker, and loss of productivity during the training period. (See Vermeulen and Hudson-Wilson 1981.) Whether these costs occur because participants perform poorly or because they quit is also valuable to explore.

\textbf{Outcomes for Participating Employers}

An employee's productivity and tractability during the training contract represent major costs or benefits to employers who take part in the JTPA program. After the contract, they may become termination employers by hiring the participant they trained. At this point all the benefits discussed above apply. Beyond these, the most obvious benefit is the \textit{subsidy to participant wages}. 
By far the most common reason employers give for participating in OJT is the subsidy. Other explanations commonly reported include eliminating the need to screen large numbers of applicants, the ability to expand or to stabilize without mounting the full cost for the new employee, and satisfaction at being able to assist deserving individuals (Simpson 1984b). Commonly reported costs include the time and supervision required to train, the potential of greater-than-average worktime lost to personal or family problems, the possibility that maximum performance after training will not match that of other employees, and the possibility that JTPA employees might turn over faster than others would.

One element of employer costs has declined dramatically since early CETA programs: the degree of constraint experienced by the employer. O’Neil’s (1982) analysis of employer hesitance to use “targeted jobs tax credits” demonstrates that the sheer fact of being constrained can be costly to employers. Earlier CETA programs protected their right to serve participants with greatest need, but in so doing, raised the employer constraint expenses above the threshold allowing participation.

Some of these costs or benefits have break-even points of zero. For example, the OJT wage subsidy cannot be costly in and of itself, and paperwork requirements cannot be seen as benefits; they can at best pose zero cost. Other benefits and costs to participating employers are meaningful only when a break-even point is defined in comparison to typical employees who would be hired were it not for the JTPA program. The two major outlays JTPA wishes participating employers to accept are hiring individuals who appear to be less qualified for the job than typical non-JTPA hires, and providing extra training beyond that required by typical non-JTPA hires. The issue is not, for example, whether the OJT employer loses five or 10 weeks of productive time during training, but whether the difference in training time for typical non-OJT hires vs. the OJT hire is zero, five, or 10 weeks. A difference of zero weeks represents a break-even point on that particular measure.

There is no a priori method to establish a balance between major costs and benefits for participating employers. Program policymakers must decide whether they are satisfied with the differences employers report
between JTPA participants and other hires, given the JTPA reimbursement they receive. Data on employer benefits and costs simply make decisions such as setting the level of OJT reimbursement more rational.

Even so, such interpretations may be less obvious than many program managers suppose. In particular, it is not always the case that employer-based programs work best for participants when employers receive maximum benefit from participation, as demonstrated by economic theory regarding nonsubsidized on-the-job training (Maranto and Rodgers 1984; Hoffman 1981). Employers always engage in introductory OJT, specific to the firm and to the job. This training represents part of the employer’s investment in hiring any new employee. The typical sequence is hire with intention to retain, invest in training, and retain as planned.

The subsidized OJT situation differs from this typical sequence in that the training occurs before the decision to retain, and the training may not be the result of a decision to invest in training. If the total cost of training a JTPA participant is greater than the income derived from the wage subsidy, the employer must decide to invest in training, which in turn implies a commitment to hire if possible, so as not to waste the investment. If, however, an SDA offers subsidies equal to or larger than the employer’s cost, the employer may participate without ever having decided to invest in the participant. The reason may be kind—”Now I can afford to help this person.” Or, it may be hard-nosed—”I make more money hiring OJTs, even if I increase turnover by letting them go after the contract ends.”

At the extreme, some participating employers use the federal wage subsidy without incurring the expense of providing any services. In-depth interviews with CETA OJT employers located some who explicitly stated that they provided no training and refused to alter their hiring practices at all, choosing instead simply to gather the windfall wherever one of their new hires happened to be OJT eligible (Simpson 1984b). Therefore, service providers are presumably well-advised to balance costs and benefits for participating employers in such a way that outcomes are positive, but not so positive as to protect employers from making an investment in each participant they train.
Outcomes Measured through Agency Records

The most basic benefits accruing to participating employers are financial, and may be recorded directly from JTPA contracts. This form of measurement is preferred because it is highly reliable, it indicates both planned and actual expenditures, and it avoids the awkwardness of asking about money during telephone interviews. In addition, agencies may be able to estimate the amount of screening and referral time they provided, thereby offsetting employer hiring costs. The question of how effective the screening was is separate, and must be measured during employer interviews.

Outcomes Measured through Employer Surveys

In addition to measures listed earlier for termination employers, participating employers incur a number of costs during their contracts, and also experience the potential costs and benefits of working with JTPA agencies. These may be measured through follow-ups, in the form of employer reports of their activities or perceptions of JTPA. Presumably the most basic costs incurred in the case of OJT are training costs. Although small employers seldom estimate training costs, they can report length of typical training for a given position, length and intensity of JTPA training compared to non-JTPA training, whether specialized methods, curricula, or personnel are used during training, etc.

Measures of perceived participation risk. Participating employers face the costs of accepting risk or constraint from their involvement with JTPA. Although particular JTPA participants may prove to be ideal workers, a program offering subsidies in exchange for hiring particular individuals has some implied risk. Employers may fear that the employee could be a poor worker, an alcoholic, or a thief. This felt risk may loom larger than the actual costs experienced when a particular worker performs poorly. The JTPA agency could also attempt to constrain the employer’s behavior, or unexpected paperwork demands could develop. These possibilities may be expensive in employers’ perceptions.

At the other extreme, the employer could reduce risk by retaining control over the hiring process. In the most extreme case, employers make firm hiring decisions before sending their new employees to apply
for the OJT subsidy. Since this practice undercuts the hiring-incentive role of OJT wage reimbursement, employer reports may be somewhat biased, depending on guarantees of confidentiality. At a more intermediate level, JTPA staff, recognizing employers’ fear of risk, may screen so carefully that the risk factor is neutralized. Thus, valuable knowledge may be gained from measures of employers’ “felt-risk,” employers’ control over screening, and the degree to which JTPA staff screen out participants least job-ready (and therefore most in need).

*Measures of employers’ direct assessment of participation costs and benefits.* Employers may also be asked for their direct assessments of the costs or benefits of participating. Most of the measures suggested thus far have been indirect, in that they ask employers to rate a particular JTPA participant or placement experience. This approach has the value of defusing employer concerns about being evaluated, i.e., clarifying that it is the employee who is being evaluated, and it allows aggregation of quite precise information regarding a representative sample of participants.

Certain employer outcomes, moreover, are best estimated in a direct form. Employers can be asked to evaluate JTPA services and staff, and to indicate how beneficial or costly they found specific aspects of participation to be. Some measures, such as the subsidy to wages, help in enlarging or stabilizing the work force, or the good feeling of helping others, can logically represent only some degree of benefit. Others, such as JTPA applicants’ screening, may represent costs or benefits, depending on their quality.17

*The Characterization of the Employment Establishment and of the Participant’s Position within It*

When one goal of employer surveys is to perform differential impact analysis of participant outcomes, employment-establishment characteristics and participant-selection-and-training characteristics should be included among employer measures. The following three levels of measurement specificity are encountered:

1. Measures characterizing the entire employment establishment, such as number of employees, industrial sector, or referral patterns established with JTPA.
2. Measures applying to any employee with the same job held by the JTPA participant, such as job complexity, qualifications required for that job, or training level of typical non-JTPA hires.

3. Measures applying specifically to each JTPA participant, such as the length of training received, or employer's ratings of that participant.

In cases where SDAs envision repeated local employer follow-ups, efficiency can be increased and nuisance to employers decreased by treating categories 1 and 2 as once-only measures analogous to those for service providers. Category 1 measures could be taken during an initial work-up with each new employer. Category 2 measures would be gathered once for each separate job title into which each employer accepted JTPA participants. Such measures are easily integrated into program operation where employer or participant analyses are envisioned. They can be combined with participant data for analysis as long as both employer and participant data include an identifier for each employer.

**An Illustration of Differential Impact Analysis Including Employer Data**

A summary of selected findings from the Washington State CETA OJT study discussed earlier will serve to illustrate the application of differential impact analysis (Simpson 1984a). That study analyzed data from a nine-month follow-up of 881 OJT participants and 517 OJT employers who trained them. In addition, data from participant MIS files, state labor and industry sources, and surveys with all OJT service providers in the CETA system were combined with data from the two follow-up surveys. Selected findings relevant to one program development issue—quality control over OJT placements—are summarized here.

There is a continuing question concerning the extent to which OJT represents a training intervention, with employers reimbursed for additional training costs demanded by their program participation, or a hiring incentive program in which employers provide little service except to hire from the list of eligibles. In addition, during the period immediately preceding the 1982-83 data collection for this analysis, the State of
Washington had decided to rather dramatically expand its OJT program, raising the question of whether pressure to enlarge the pool of new employers led to a deterioration in the quality of OJT placements.

While gross impact analysis cannot assess the net impact of OJT for participants, we were able to test a series of propositions comparing forms of OJT implementation that placed greater or lesser emphasis on quality control. We were also to compare specific OJT placements that appeared to provide service to participants of greater or lesser quality. The following summarizes findings from differential impact analysis of these implementation and treatment variants on postprogram outcomes. All of these reports are for the outcome variable most clearly affected by programmatic variables—whether or not one retained employment with the original OJT employer. 18

1. Various qualities of OJT implementation and treatment explain far more variation in outcomes than do the full set of participant background characteristics included in MIS files and augmented by measures on the follow-up survey, although age and employment history have considerable impact. This was true in part, we learned, because so many OJT positions were entry level, making such minimal demands that some highly qualified OJT enrollees left voluntarily, thus undermining program success.

2. Service provider measures indicating the degree to which strong quality control procedures were a part of their OJT program implementation proved strongly associated with the rate at which participants retained their OJT positions through the nine-month follow-up. A number of factors raised OJT retention: more demanding quality review for new and old employers, a policy demanding higher than minimum wage for OJT placements, and a willingness to hold some money back because an insufficient number of satisfactory employers were available for OJT. We estimated a 28 percent difference in OJT retention rates, above and beyond other factors, between agencies placing most and least emphasis on quality control.

3. Consistent with the interpretation that OJT in this system was suffering from low quality control, a set of measures designed to measure the quality of participant training also indicated higher OJT
retention where training had been more intensive. In particular, employers who reported offering any special instruction for OJT participants, who used more formal preparation along with informal OJT, and who had more personal involvement in the training retained their participants more often. The impact of these factors was modest because training intensity overlaps with two other powerful predictors of retention: the complexity of the job, and the participant’s enthusiasm for the work.

4. While it is common to worry about OJT participants’ ability to meet their job demands, the reverse proved much more problematic in this study. The less background the participant reported having in the area of the OJT job and the more complex the employer described the job as being, the higher the OJT retention (after adjustments for participant background characteristics). We found that only 7 percent of the participants were fired for inability to do the work, while 10.3 percent gave boredom with the job or getting no training as the main reason they quit, and another 14 percent left for a better job. In all, 31 percent quit, while 21.7 percent were fired or left by mutual agreement with the employer.19

5. The three strongest predictors of OJT retention were the employer’s rating of how enthusiastic and cooperative the participant was, the employer’s rating of how fast the participant worked, and the participant’s felt importance of retaining a career in the type of work represented by the OJT job. All these turn out to be much higher when the job is more complex, when OJT positions provide more training, when participants are moving into a new area of work rather than being placed in a job about which they know a great deal, when employers more frequently provide evaluative feedback to participants, and where service providers emphasize quality control. The higher the quality of the OJT placement, the more likely participants were to like the job and treat it as a career they valued, and in turn display behaviors employers wanted to see.

These findings conclude that in that particular OJT system at that time the program needed to develop quality control over the nature of the OJT site, i.e., the services offered by employers. One other troublesome
finding consistent with this concern for quality was the discovery that
some employers in that sample were explicitly using the OJT reimburse-
ment as windfall profit.

Although employer-initiated OJT s were quite common in this sample
(45.8 percent of all OJT s), most of these represented referrals by
knowledgeable employers who made no hiring decision until eligibility
was established. Postprogram success in these cases was no higher than
average. However, one-sixth (17.3 percent) of the employers we inter-
viewed said they first made a firm hiring decision and then sent the
participant to CETA to see if a wage subsidy could be gained. 20 This
phenomenon represents both a poor expenditure of training dollars and
a selection mechanism likely to bias outcome estimates. Among these
participants, retention was 12 percent higher, after adjusting for other
factors.

This set of findings was chosen to illustrate the value of differential
impact analysis because many separate tests lead consistently to the same
conclusion, and because that conclusion is in essence opposite to the
normal interpretation of weak program performance. When one service
provider performs at a higher (descriptive) rate than another, nearly all
analysts will ask whether that difference was produced by “creaming,”
i.e., whether the finding represents selection bias. Few will ask, however,
whether participants were too highly qualified, relative to the quality of
the OJT jobs and training. Yet, careful quantitative analysis of program
implementation and treatment confirms the latter interpretation in this
one service delivery system.

How Gross Impact Analysis Compplements Net Impact and Process Evaluations

The most valuable uses of the gross impact evaluation method—and
also its major limitations—may be placed in relief by a brief examination
of the ways in which the three approaches in this volume complement
each other. The gross impact approach exploits its measurement flexibil-
ity to enlarge the range of outcomes analyzed as well as the range of
factors considered as influences on outcomes, yet the quantitative nature
of its measures helps insure that conclusions are reliable.
In the Washington State CETA OJT example, the gross impact analysis provided no knowledge of the net value of that OJT system to its participants. If we had employed both the net and the gross impact approaches, the major value of the gross impact findings would have been to broaden the range of measures analyzed—both outcomes and programmatic factors that might have influenced the net impact. These measures would have increased our ability to explain how the system works to produce the high or low net impacts we identified, providing guidance on improving net impact. Without a net impact evaluation, we do not know how urgent the need for system improvement is. Without the gross impact evaluation, we have less guidance regarding the mechanisms needed to improve a system.  

At the other end of the continuum lies process analysis. Its detailed analysis of program implementation feeds gross impact analysis by identifying measures worthy of quantification. Only when the analyst understands the process by which organizations operate will meaningful outcomes be selected for measurement, or will the analysis of program-variant effects upon outcomes be meaningful. The centrality of implementation factors to differential impact analysis means that gross impact analysis is in part quantified process analysis and should always be preceded by at least a partial process analysis.

In addition to its focus on outcomes, gross impact analysis adds four major complements to process analysis: (1) *postprogram* outcome measures, (2) quantitative precision of measurement, which allows reliable estimates of the impact of program variants on outcomes, (3) reliable comparisons across multiple service providers or treatments, and (4) measurement of *individual service treatments* as possible influences on outcomes, along with measurement of the implementation factors also emphasized by process analysis.

Because process evaluation avoids the limits of formal, quantitative data collection, it can be flexible, creative, and unique. However, for the same reason, i.e., process analysis does not collect quantitative data, conclusions from process analysis are subject to considerable error, which can be reduced by subjecting the conclusions of process evaluations to a gross impact analysis. Such interpretations can be tested in
multiple contexts and can include a range of measures indicating individual treatment as well as program implementation. Thus, for example, a process evaluation conclusion that an organization is less effective than it might be because of some particular element in its internal structure can be tested by comparing outcomes, as indicators of effectiveness, among organizations differing with regard to that structure.

The CETA OJT study, discussed in order to illustrate differential impact analysis, began its conceptualization phase with an informal process analysis. Our aim was to identify alternative theories regarding OJT system operation, program variants likely to impact outcomes, and program variants that could be changed if policymakers decided to use our findings to improve program performance. While standard variables for such research were also measured, several analyses that proved to be most fascinating involved variables that emerged from the process analysis. Had the research ended at the process analysis stage, these ideas could not have been tested and could not have generated quantitative estimates of the impact which program variants can have on outcomes.

Chapter Summary

The major goal of the gross impact evaluation approach presented here has been to improve the technology available to managers of human services organizations. Technology, in this case, is knowing how to operate programs that effectively produce the desired outcomes—transforming clients with given needs at intake into postprogram success stories. Gross impact analysis approaches that goal with two distinct analytic strategies: the analysis of descriptive gross outcomes, and differential impact analysis. These analyses are performed using data from several possible sources: MIS files, participant interviews, data from service providers on individual treatments and program implementation, and data from others, such as employers, who may be closely involved in the operation or outcomes of the program.

The analysis of descriptive gross outcomes is useful because it is simple. It is also dangerous, for the same reason. The following steps may be taken to enhance the usefulness of descriptive outcome measures: (1)
broadening the range of outcomes studied, to provide fuller interpretations of how the program is functioning; (2) tailoring measures and analysis questions to make findings more meaningful; (3) measuring views of those whose perceptions have *prima facia* meaning (e.g., employers); and (4) insuring high technical standards during data collection and analysis. Analysts can also protect against misuse of descriptive reports by limiting the nature of their interpretations to those merited, given the limitations of the analysis.

The second analytic approach, differential impact analysis, is more expensive and also more useful for program development. This method increases the level of technology available to managers by testing the relative effectiveness of alternative forms of program implementation or service treatments provided to individuals. That is, differential impact analysis estimates the degree to which difference in postprogram outcomes is caused by any given program variant. Each alternative is tested against other alternatives which are in place in the service delivery system.

Further, differential impact analysis attacks the problem of selection bias, the primary factor limiting the validity of descriptive gross outcome reports and inhibiting causal interpretations. By using a variety of sources to measure selection processes as well as client characteristics predictive of postprogram success, the most powerful "alternative explanation" facing all program evaluation—that outcomes were produced by client characteristics or selection rather than by the program—can be greatly mitigated, if never completely eliminated.

Thoughtful preparation of survey data collection tools and appropriate use of analysis techniques, available in a wide range of statistical packages, can make differential impact analysis a powerful tool for improving program performance on mandated outcomes. At the same time, the wide spectrum of measurable and describable gross outcomes can be used to improve or maintain the quality of service while core outcomes are being maximized. Findings apply directly to the state or local service delivery systems in which the evaluation was conducted, directing managers to increase some services or retain some implementation forms, while reducing others. Local applicability of findings,
protection against selection bias, and testing of alternative program variants provide the basic components of quasi-experimental causal analysis, so that careful differential impact analysis produces results that can more accurately identify factors influencing program effectiveness. Changing these factors is, therefore, likely to produce improvement in effectiveness.

NOTES

1. Whether each group is better off because of participating in the program is a net impact question. Gross impact analysis identifies which services work better for each group.

2. The range of available alternative explanations may be identified by asking managers of programs that are performing poorly to explain their organizations’ weak showing.

3. The aim of all research is to increase or reduce our confidence in particular conclusions or interpretations. To reject new information because it is imperfect is as foolish as to embrace unreliable findings wholeheartedly. The analyst must assess the value of any research finding. Structuring the research so that findings are firmer improves the value of the research, even if the method remains considerably less than perfect.

4. Since these results are descriptive only, this interpretation is also subject to error. It could be, for example, that more AFDC recipients leave their OJT positions because they are unstable, while non-AFDC recipients leave in order to move to higher paying jobs in nonrelated areas. Such possibilities can be tested if the researcher has entertained them early enough to make the data available. In this case, for example, more non-AFDC people did quit for better jobs, but not enough more to invalidate the initial interpretation offered in the text.

5. This figure and others in this series may be somewhat lower than typical since Washington was experiencing rather serious recession during this study.

6. Above and beyond the obvious reason that JTPA now requires such a definition.

7. One topic of particular interest to small SDAs involves the correction downward of the sample needed for a given error margin when the population from which the sample is drawn is very small.

8. This does suggest one possible pitfall of low-budget program evaluation: if surveys are conducted in-house, respondents may bias their answers in a positive direction. Sophisticated external consumers will therefore tend to question findings based on surveys, unless they are conducted by third parties and guarantee confidentiality.

9. These membership identifiers also serve as extremely important control variables under conditions discussed later in the chapter.

10. The term “membership identifier” has been used throughout this chapter to refer to what statisticians typically call “dummy variables.” The standard rules governing proper analysis of dummy variables should be followed during the analysis described here. If all membership identifiers are entered simultaneously, only n-1 may be included. For example, if 15 service providers are included in the sample, membership identifiers (dummy variables) indicating 14 may be included in a single regression equation. If the analysis involves a forward stepwise procedure, n (all) membership identifiers may be included.
11. It should be noted that because interaction terms are highly correlated with their constituent variables, analysts should consult the change in variance explained ($R^2$) rather than relying on $t$ or $F$ scores for each individual variable within the model.

12. When programs are most successful, nearly all participants are employed after the program. Variance in earnings is greatest when a large proportion of participants earns nothing.

13. Optionally, a written form with anonymous return could be administered by staff, but not without the problems of return rate and secretarial overload which accompany mail surveys.

14. Examples of such questions as they appear in a survey are available in Simpson (1986).

15. For a set of specific suggested measures that have been pre-tested among JTPA employers, see Simpson, 1986.

16. One difficulty with survey data emerges here. Respondents often tend to bias their reports in a positive direction, so that the midpoint of any set of answers is always a bit above the face validity midpoint. Thus, if a group of employers were asked to rate all their employees, the average employee would be rated somewhat above average. Analysts should bear this in mind when interpreting employer ratings. However, no precise information exists with which to estimate to what extent responses are inflated, making adjustments imprecise.

17. See Simpson (1986) for specific measurement suggestions for this issue as for other employer issues.

18. One of our more intriguing findings was that the ability to regain other employment once the OJT position was lost was almost entirely immune to interpretation via program variables. That is, the impact of the OJT program—or, rather, any variations in its implementation—extended only to getting and retaining the original OJT job.

19. For valid measurement of issues such as this one, it proved especially useful to have data from both the employer and the participant.

20. Presumably, employers were so candid with us because we were a neutral third party. Some employers even offered explicit statements that they viewed the entire process, cynically, as a windfall.

21. It is possible in theory, and at great expense, to conduct net impact evaluations that include a broad range of outcome measures. The cost-efficient design suggested in this volume by Johnson foregoes this possibility to make the research feasible for states and large SDAs. Any quantitative analysis of a wide range of program variants automatically becomes a gross impact analysis because these measures are meaningful only among participants. An untreated comparison group must, therefore, be omitted from such an analysis.
REFERENCES


