USING THEORY TO IMPROVE PROGRAM AND POLICY EVALUATIONS

Edited by Huey-tsyh Chen and Peter H. Rossi

Prepared under the auspices of the Policy Studies Organization
Stuart S. Nagel, Series Adviser

Contributions in Political Science, Number 290

Greenwood Press
New York · Westport, Connecticut · London

Contents

Introduction: Integrating Theory into Evaluation Practice 1
Huey-tsyh Chen and Peter H. Rossi

Part I. Concepts and Nature of Theory-Driven Evaluations 13

1. Four Types of Theory That Can Guide Treatment Evaluations 15
John W. Finney and Rudolf H. Moos

2. Theory-Driven Metaevaluation 29
William R. Shadish, Jr.

Part II. Research Strategies and Methods 47

3. Pattern Matching in Theory-Driven Evaluation: A Field Example from Psychiatric Rehabilitation 49
William M. K. Trochim and Judith A. Cook

4. Testing Theories in Theory-Driven Evaluations: (Tests of) Moderation in All Things 71
Melvin M. Mark, David A. Hofmann, and Charles S. Reichardt

5. Theory-Driven Meta-analysis: Practices and Prospects 85
David S. Cordray

6. The Role of Theory in Doing Case Study Research and Evaluations 97
Robert K. Yin

7. Theory in Evaluation: We Think, Therefore We Theorize (An Ethnographer's Perspective) 115
David M. Fetterman

Part III. Applications in Program/Policy Planning and Implementation 127

8. Using Research and Theory in Developing Innovative Programs for Homeless Individuals 129
Debra J. Rog and Robert B. Huebner

9. Evaluating New Dawn and Pegasus Using the Chen and Rossi Multigoal, Theory-Driven Approach 145
Dennis J. Palumbo and Peter R. Gregware

10. Using Program Theory in Quality Assessments of Children's Mental Health Services 165
Keith A. Peterson and Leonard Bickman

Part IV. Applications in Impact Assessment and Cost-Benefit Analysis 177

11. The Importance of Theory in Selection Modeling: Incorrect Assumptions Mean Biased Results 179
David Rindskopf

12. Quasi-experiments and Causation Probing 193
Huey-tsyh Chen and Lung-ho Lin

13. Research-Based Theory for Educational Planning and Evaluation 207
Herbert J. Walberg and Arthur J. Reynolds

14. Meta-analysis in Evaluation Research: Moving from Description to Explanation 229
Mark W. Lipsey

15. Theory-Driven Approaches to Benefit-Cost Analysis: Implications of Program Theory 243
Anne Scott and Lee Sechrest

Select Bibliography 259

Name Index 263

Subject Index 269

About the Editors and Contributors 273

Chapter 3

Pattern Matching in Theory-Driven Evaluation: A Field Example from Psychiatric Rehabilitation

William M. K. Trochim and Judith A. Cook

Patterns are the crucial link between theory and data. A theory describes what we believe happens, and perhaps why. It can consist of rough guesses or hunches, or it can be delineated formally or mathematically. Data depict some aspect of what is actually happening in reality. Data can range from informal observations or recollections, to a multivariate quantitative measurement structure. In order to see whether our theories make sense, we must put them up against data to look for a correspondence. Patterns are the forms we use to represent both theories and data in order to treat them in comparable terms. In this sense, patterns are a translation device, the common ground where theories and data, ideas and observations, can meet.

Patterns occur in all aspects of the research endeavor, in both the theoretical and observational realms. Very often in applied social research, the patterns are simple binary ones--if we do *x*, then *y* will occur, and if we do not do *x*, then *y* will not occur; if we administer a program designed to help children in mathematics, then they will perform better on mathematical tasks, and if we do not, they will not improve. When we find the predicted pattern in the observed data, we can conclude that our theory *may* be correct. It has certainly not been refuted. But the problem with simple binary patterns is that they are simple –there are usually too many theories other than our own that would predict the same pattern in the data. Thus, if we find such a binary pattern in the data, we need to demonstrate that it is there because of our theoretical expectation, and not for other reasons.

H. Chen and P. H. Rossi (1980, 1983, 1987) have reminded us all of the fragility of this binary pattern matching model. They have correctly warned that we must break open the "black box" of this binary expectation pattern and generate theoretical patterns that are far more specific descriptors of reality. One

can read their work as a call for more complexity and specificity of theoretical patterns. The advantages are readily apparent, and this is what makes their call for "theory-driven" evaluation so compelling. For instance, assume that we go inside the black box of a program or independent variable and, instead of treating it as an on-off switch, we measure the *amount* of treatment that is actually administered to each person. Now, our theoretical pattern would consist of the variable amounts of treatment, and our theory would predict that more treatment would be associated with greater gains in performance of treatment-related tasks. If we find a correspondence between these patterns, can we conclude that our theory is correct? Of course not. There may, as always, be other reasons that we have observed a relationship between higher levels of treatment and improved treatment-related performance. But something is certainly different between these two examples—the simple binary case and the more differentiated treatment levels approach.

A fundamental principle implicit in pattern matching is that more complex theories generally have fewer competitors—there will be less likelihood that there are other suitable explanations for the more complex expectation pattern. In a sense, that is the good news. The bad news is that it is probably more difficult to find a match to a more complex specific pattern. With more specific patterns, there are more ways that we could be wrong by chance alone or as a result of errors in measurement. But if we do find a match—if we find better treatment-related performance with higher levels of treatment—we can probably be more confident in our theory than we could as a result of a binary, "black box" test. Pattern matching emphasizes the importance of uniqueness in the theoretical patterns that we generate. To be tested well, our theory should be like a fingerprint, distinguishable from all others. Our data must also be collected in a manner that allows that fingerprint to be seen. This is akin to the idea of a critical test as described in theory-building contexts in the natural sciences.

The move toward theory-driven evaluation is in the right direction, but it is only a beginning. Most writers in this area have emphasized the need for *program theory* (including contextual factors). Still, little attention has been given to developing theories related to the measurement constructs that will reflect reality or to devising theories of the participants and how they vary. Moreover, we have only begun to think about how we might link complex program theory models with complex multivariate measurement structures and participant theories. Likewise, we have only begun to explore the different *forms* that our theories might take. Consequently, our ideas regarding the types of patterns that might be involved in theory-driven research are primitive. We certainly have no classification of pattern types that approaches being definitive.

This paper is a call for greater attention to the role of pattern matching in theory-driven research, especially in applied social research and program evaluation contexts. The recent literature on theory-driven research is briefly reviewed to discover what different theory forms are being discussed. The general idea of pattern matching is then presented, and its relationship to theory is explored.

Finally, a brief example of a real pattern matching study is offered, along with a discussion of some of the methodological issues that it raises.

THEORY IN EVALUATION

In terms of program evaluation, the 1980s can perhaps best be characterized by the rediscovery and reemphasis of the role of theory. Chen and Rossi's work on theory-driven evaluation (1980, 1983, 1987) spans this period. Others have added to this reemphasis on theory, leading to several excellent edited volumes (Bickman, 1987, 1990; Chen, 1989a) and a book-length treatment of the topic (Chen, 1990).

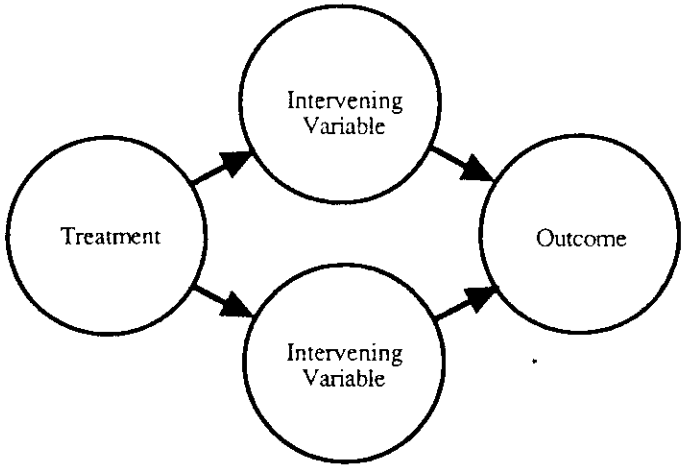
What do we mean when we talk about "theory"? What is a "theory," and what does it look like? Various theory types have been suggested in the literature on theory-driven evaluation. The most common is some form of a *causal model*. Causal models are usually depicted in graph form (although they can also be portrayed as a set of equations). The simplest causal model graph would show a treatment and an outcome and join them by an arrow that implies causal direction. This is illustrated in figure 3.1.

This is, in effect, the simple black box model portrayed in causal map form. It might be tested using a t-test or a one-way ANOVA. The treatment could be represented as a binary number, a 0 indicating absence and a 1 indicating presence. The outcome could be a continuous quantitative measure of a key dependent variable of interest. The theory-driven evaluator would reject this model as too simplistic. There are many ways in which they might like to "complicate" it. Chen (1989b, 1990) and Chen and Rossi (1989), for instance, have argued for the inclusion of intervening variables. Figure 3.2 shows one possible graphic depiction of such a model.

M. W. Lipsey and J. A. Pollard (1989) have described separately the "basic two-step" version, which includes only a single intervening variable; they have claimed that it is "the most common theory form in contemporary evaluation research" and that "it is the easiest to incorporate in practical program evaluation" (p. 320).

Figure 3.1.
A Simple Causal Model

Figure 3.2.
A Causal Model with Intervening Variables

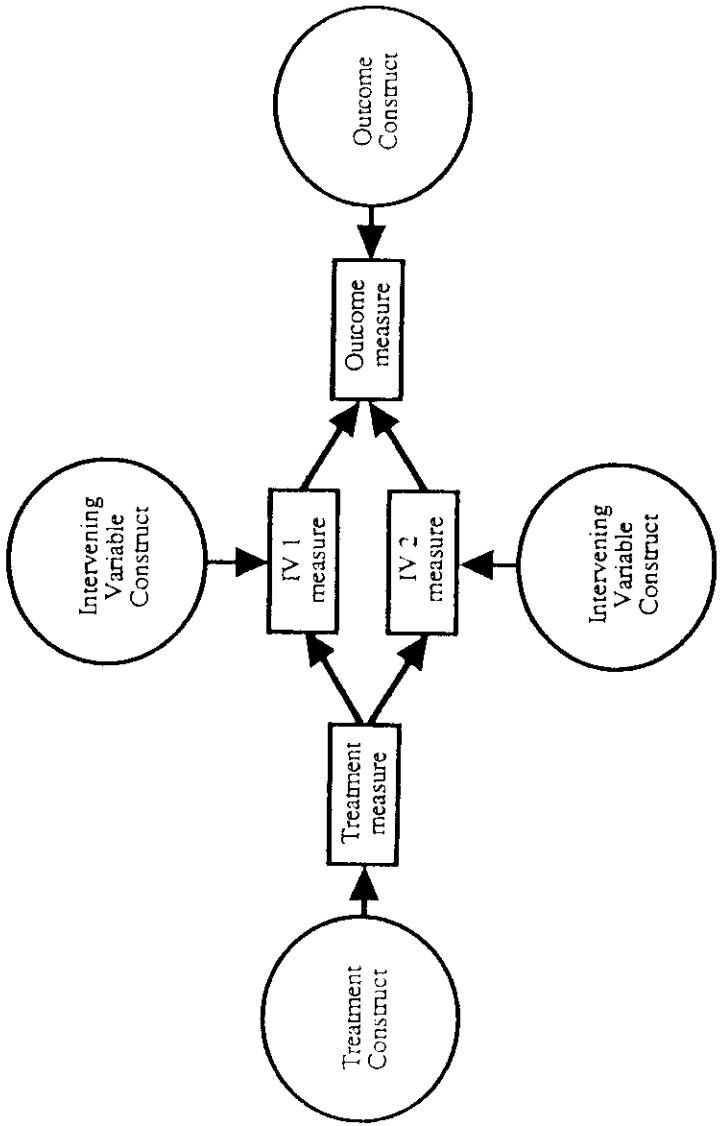


H. L. Costner (1989), Lipsey and Pollard (1989), and undoubtedly many others follow the tradition of structural equation modeling by making distinctions between unmeasured constructs and measured variables. This more complex version can be depicted as in figure 3.3, where constructs are indicated by circles and their corresponding operationalizations are enclosed in rectangles. To this, we can add separate symbols for measurement errors, extraneous factors, contextual factors, and prior demographic and descriptive covariates, all of which may be joined by causal or relational arrowed lines as our theory might dictate (including feedback loops as discussed in McClintock, 1987), yielding a quite complex figure indeed.

All of these represent the *causal model* approach to theory-driven evaluation—undoubtedly the most common representational form that has been discussed in the literature. We might also include in this camp the discussions of J. W. Finney and R. H. Moos (1989; see also their figure 1.1 in this volume) who use symbols to distinguish between life context factors prior to treatment, client pretreatment factors, intervention factors, life context factors following treatment, and client posttreatment factors. Causal models are in the mainstream of social research methodologies and are intimately related to structural equation models, regression analysis, and path modeling.

Despite the dominance of the causal model perspective, there have been other suggestions. Lipsey and Pollard (1989) have described work by W. K. Runyan (1980) on stage-state models that might be appropriate for theory-driven evaluation. Here, the key stages that program participants can pass through are de-

Figure 3.3.
A Causal Model with Intervening Variables Which Distinguishes Constructs from Measures



scribed along with the "states" or statuses that people can take within each stage. Frequencies are then compiled for which states participants are in at each stage. From these, it is possible to depict the most common sequences through the stages and also to examine interesting outlier cases that might suggest where treatment goes wrong or interacts with other factors of importance.

We might also distinguish what can be termed "condition-based" models. Simply put, a number of factors deemed relevant – contextual, programmatic, person-based—can be described and measured, and their relationship with outcomes can thus be examined. The chief difference between this model and the causal model approach is that there is no attempt to develop a sequential or path-oriented depiction. All of the conditions are treated as potential correlates of outcome, but no attempt is made to treat some of them as "intervening" or coming between the treatment and outcome. We might read A. G. Scott and L. Sechrest (1989) as falling partially into this camp. They emphasize a number of conditions or factors of relevance in a treatment, such as its purity, specificity, dose, intensity, and duration. Presumably one could correlate each of these (or some linear composite of the set) with outcome variables to examine effects. Also falling into this approach would be factor analytic models and facet theory approaches.

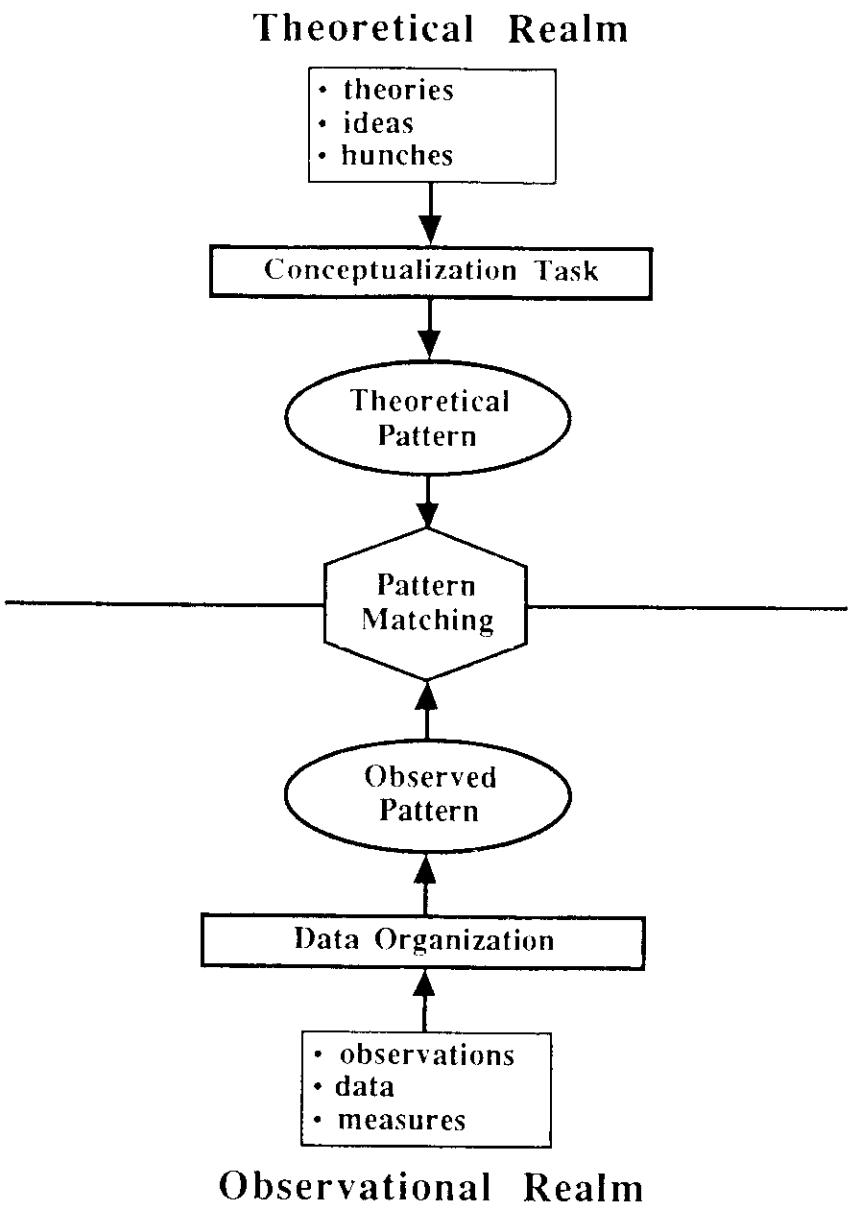
These three models—causal, stage-state, and condition-based – are certainly not an exhaustive list of the types of models that might be available to the theory-driven evaluator, but they are indicative of the major theory representation forms that have been discussed. Clearly, the field is dominated by a causal modeling approach, and it is this type of model that will be discussed later in connection with pattern matching.

THE IDEA OF PATTERN MATCHING

A pattern is any arrangement of objects or entities. The term "arrangement" is used here to indicate that a pattern is by definition nonrandom and at least potentially describable. All theories imply some pattern, but *theories and patterns are not the same thing*. In general, a theory postulates structural relationships between key constructs, as shown in the causal modeling approach described above. The theory *can be used as the basis for generating patterns of predictions*. For instance, $E = MC^2$ can be considered a theoretical formulation. A pattern of expectations can be developed from this formula by generating predicted values for one of these variables, given fixed values of the others. Not all theories are stated in mathematical form, especially in applied social research, but all useful theories provide information that enables the generation of patterns of predictions.

Pattern matching always involves an attempt to link two patterns, of which one is a theoretical pattern and the other is an observed or operational one. The basic idea of pattern matching is illustrated in figure 3.4.

Figure 3.4.
The General Pattern Matching Model



The top part of the figure shows the realm of theory. The theory might originate from a formal tradition of theorizing, might be the ideas or "hunches" of the investigator, or might arise from some combination of these. The conceptualization task involves the translation of these ideas into a specifiable theoretical pattern (indicated by the upper oval in the figure). The bottom part of the figure indicates the realm of observation. This is broadly meant to include direct observation in the form of impressions, field notes, and the like, as well as more formal objective measures. The collection or organization of relevant operationalizations (i.e., relevant to the theoretical pattern) is termed the observational pattern and is indicated by the lower oval in the figure. The inferential task involves the attempt to relate, link, or match these two patterns as indicated by the double arrow in the center of the figure. To the extent that the patterns match, one can conclude that the theory, and any other theories that might predict the same observed pattern, receive support.

It is important to demonstrate that there are no plausible alternative theories that account for the observed pattern, and this task is made much easier when the theoretical pattern of interest is a unique one. In effect, a more complex theoretical pattern is like a unique fingerprint that one is seeking in the observed pattern. With more complex theoretical patterns, it is usually more difficult to construe sensible alternative patterns that would predict the same result. To make this more concrete, consider a theoretical outcome pattern for an educational program evaluation; it is hypothesized that the program will have its greatest effect on measures of immediate recall of course-related information, less of an effect on attitudes, and the smallest effect on behavioral changes. If this pattern of outcomes is obtained, one might be tempted to conclude that the program caused the observed pattern to occur. There may, however, be a plausible alternative explanation for this observed pattern. For instance, it may be that recall measures are more reliable than attitudinal ones, which in turn are more reliable than behavioral ones. The observed pattern in this instance may be due to the pattern of reliability across measures rather than to the program. In this case, one would have to rule out the reliability-based explanation (perhaps by examining reliabilities to see if they are in fact distributed in this manner or by incorporating more measures of each type with differing reliabilities) before concluding that the program caused the outcome pattern. To the extent that theoretical and observed patterns do not match, the theory may be incorrect or poorly formulated, the observations may be inappropriate or inaccurate, or some combination of both states may exist.

In any program evaluation setting, there are many opportunities for pattern matching. W. Trochim (1989d) has classified the major pattern matches in program evaluations into two major types: process pattern matches and outcome pattern matches. The *process pattern matches* can be further subdivided into three types: program, measurement, and participant pattern matches. Program pattern matches compare the theory or idea of the program with its implementation and can be viewed as addressing the construct validity of the cause (Cook & Campbell, 1979). The measurement pattern matches compare the conceptual theory of the interrelatedness of the measures with their observed intercorrelations. This is in the tradition of the nomological network (Cronbach & Meehl, 1955) and the multitrait-multimethod matrix (Campbell & Fiske, 1959) approaches to the construct validity of the measures. The participant pattern match compares a theory of participant similarities with the observed demographics. This is analogous to examining the representativeness of a sample from a population. For any process pattern match, one can construct an *outcome pattern match*. When we compare our expectations of outcomes across variations of a program or in terms of different program components, we are conducting an outcome pattern match in reference to the program pattern match. Similarly, when we compare how we think different types of participants will perform with how they actually do perform, we are linking an outcome pattern match with a participant one. Finally, when we compare how we think a group of measures will be differentially affected by a program, we are combining outcome pattern matching with a measurement pattern match. It is this last example (outcome-measurement pattern matching) that may be feasible in many program evaluations and that we examine in our example.

Pattern Matching and Proximal Similarity

Pattern matching as outlined here assumes that there is a natural coherence in reality that our theory should reflect. To depict this coherence, we borrow the term "proximal similarity" from earlier work by D. T. Campbell (1986). While Campbell has suggested the idea of proximal similarity as the basis for external validity or generalizability, we believe that it is a more fundamental principle for research that has far-reaching applicability and is particularly consonant with pattern matching thinking.

Proximal similarity implies that the more similar two things are conceptually, the more similar their manifestations in reality. It assumes, as Campbell (1986, p. 74) has stated, that "nature is 'sticky,' 'viscous,' proximally autocorrelated in space, time, and probably n-dimensional attribute space, with adjacent points more similar (as a rule) than nonadjacent ones." It assumes a semantic theory that concepts differ along some dimensions of similarity and that this is so because reality consists of entities that differ along gradients of similarity.

Why is this relevant to pattern matching? In much social research, especially in applied research and program evaluation, we have little prior basis for the development of theory. We have not had great success in developing notions of fundamental social elements or units upon which our theories can rest and that are reliably measurable. In order to generate the theoretical patterns needed for pattern matching, we need some principles that enable us to state our differential predictions with reasonable specificity. The idea of proximal similarity may provide us with a general principle or "metatheory" that can help us to specify our theoretical patterns meaningfully.

How could the principle of proximal similarity be used in developing theoretical patterns? Assume that we are conducting an evaluation of a program under rather simple conditions. We have the program group, a simple comparison group (randomly assigned or not), and several measures of outcomes, each of which has multiple items. The typical theoretical pattern would be the familiar binary one discussed earlier—for each outcome, we might predict differences in average performance between treated and untreated persons. Probably no one believes that we would theoretically expect the *same* treatment effect on each of the outcome measures or even on each item within a composite multi-item measure. After all, at a semantic level, each item is distinct and must reflect slightly differing shades of meaning at least. Would we not expect that some items would reveal slightly greater treatment effects than others? On what basis would such an expectation stand? Cognitively, we might propose that some of the outcome items are "more centrally related" to the construct that we think will be most directly affected by the treatment. That is, we usually have an idea of the kinds of concepts on which our program should have its strongest effects. Assume that we can array all of the outcome items along some dimension(s) of similarity to the major proposed effect. We might then argue that treatment effects will be greater for items that are closer (more proximate) to this hypothetical major outcome. We obtain a far more specific theoretical pattern in this case because we combine our usual binary expectation with a theory of the proximal similarity among outcome items, yielding an ordered listing of expected treatment effects. At the root of this logic is the semantic assumption that meaning can be accurately construed in terms of gradations of similarity and difference (rather than as clear distinctions of class or kind).

The principle of proximal similarity can be applied to any major component of the research process—not just to the outcomes—to yield a more detailed theoretical pattern. For instance, consider the participants in a research study. No reasonable person would argue that in theory a treatment will affect all participants equally. Nevertheless, we typically average the scores of all persons in the treatment group to try to ascertain the "main effects" of the treatment. Proximal similarity suggests a way to sharpen our theoretical prediction pattern and attain greater predictive detail. The central assumption is that persons who are more alike would respond more similarly to the treatment. If we array the persons in our sample along certain similarity dimensions and describe the type of person who should be most affected by the treatment, we could predict a declining pattern of effects for persons further away on these similarity dimensions. C. M. Judd and D. A. Kenny (1981) have stated this logic well, again concentrating most on the implications for generalizing:

> We might conceive of different theoretical populations as departing from the sampled population along a gradient of similarity. Some theoretical populations are quite similar to the sampled population; others are less similar. Generally the confidence that we have in generalization to populations not operationalized depends on the population's location on this gradient of similarity. Such gradients of similarity can also be used for generalization to outcome, treatment and setting constructs that were not operationalized in the research. (p. 41)

This view of proximal similarity deliberately minimizes the traditional distinction between internal and external validity. Campbell (1986) and Judd and Kenny (1981) have described the ideas of proximal similarity and the "gradient of similarity" as the basis for *external* validity, but using proximal similarity to specify theoretical patterns helps to define a far more detailed "fingerprint" for the program. If this theoretical pattern is matched in the data, we have simultaneously enhanced *internal* validity—we have greater confidence in the causal theory that the program in question (and not some other factor) has had an effect. The traditional stance that pits internal and external validity against each other may be mistaken. Paralleling the reasoning in Trochim (1985), we can argue that *increasing theoretical specificity using proximal similarity may simultaneously enhance both internal and external validity.*

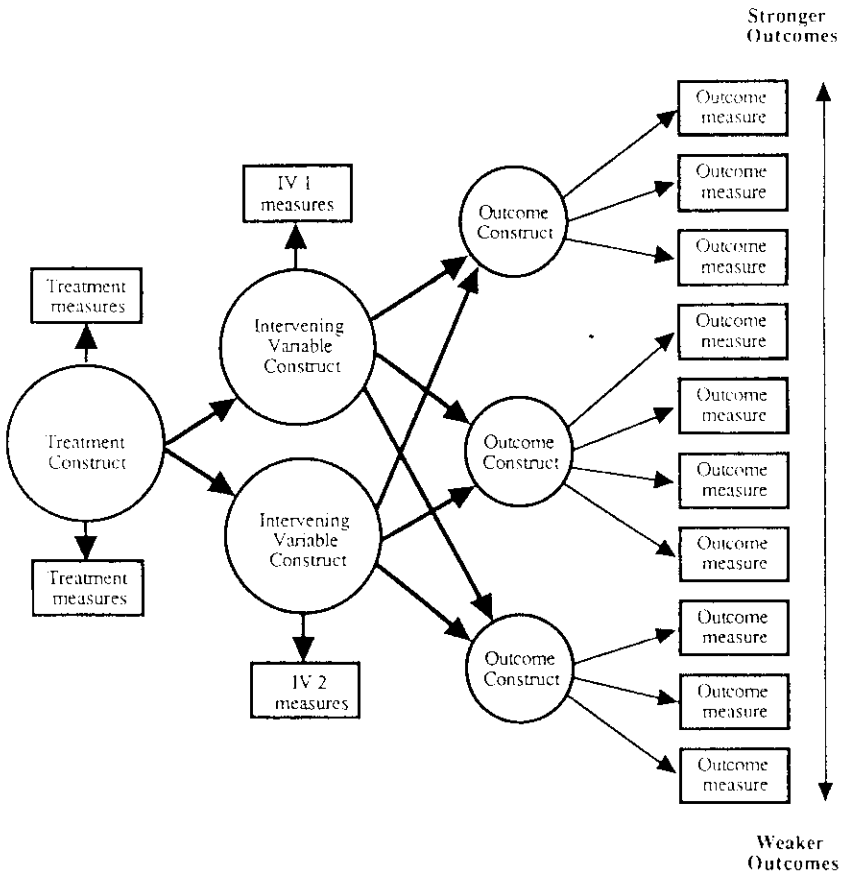
PATTERN MATCHING AND PROGRAM THEORY

Earlier it was argued that the most common form for theory-driven evaluation is a causal model. In this section, we will examine a simple variation of the typical causal model that gives it a more pattern matching flavor. Consider the model shown in figure 3.5. This is essentially a variation on the model depicted in figure 3.3 above. The major difference is in the specification of the outcomes. Here, instead of a single outcome construct or variable, the theory considers ten separate outcomes based on several different constructs. These might be ten different measures or ten items from the same multidimensional scale. In addition, the theory specifies the *ordering of the effects* for these ten outcomes. According to the theory, the top variable is expected to show the strongest effect, while the bottom variable is expected to have the weakest. Ideally, we would want to specify these hierarchical outcome expectations on an interval scale, but at least they must be ordinal.

How might this expectation pattern be generated from the causal model of figure 3.3? Three possibilities are mentioned here. First, the simplest approach would be a *direct rating method* for generating the theoretical outcome estimates. The evaluator, program administrators, or other relevant constituencies might simply examine the ten outcome measures and rate or rank order them subjectively from highest to lowest expected effects, given what is known about the program. The average ratings or rankings would constitute the theoretical expectation pattern for the outcomes. This approach is used in the example provided later.

A second strategy can be called the *proximal similarity method*. This could be accomplished in three steps. First, various stakeholders (including social science theorists) could be involved in a process to select the key outcome measure out of

Figure 3.5.
An Outcome Pattern Matching Model with Ten Outcome Measures Arrayed from Weaker to Stronger Expected Outcomes



the set of all measures. The key outcome is the measure that would be expected to be most affected by the program. For a mathematics training program, for example, this might be a recall test that covers exactly the material covered in the training. The second step would be to rate the remaining nine variables in terms of their general conceptual proximal similarity to the key outcome measure. Finally, these similarities can be scaled to yield a hierarchical outcome pattern in which the measures that are more proximally similar to the key outcome would be expected to show higher effects than those that are more distant.

Neither of these first two strategies for specifying the theoretical pattern tells

us *why* that pattern is expected. We might assume that the theoretical pattern is generated using an implicit theory, but many different theories could yield the same or similar patterns. Thus, if a pattern match is detected with one of these methods, it is important to try to elucidate the theory that persons used to generate the theoretical expectation pattern. This is not likely to be a trivial issue because persons who make predictions may not be aware of the theory that drives them or may find it difficult to articulate it. Perhaps the most straightforward approach would be to discuss the predictions with the theoreticians and, in a systematic way, ask them to describe or justify the basis for them. This would be best accomplished before looking at the observed patterns so that post hoc explanations designed to fit the facts might be avoided. The link between theory and theoretical pattern has received little methodological attention to date. Clearly, more work is needed in this area. This stage of translating from the theory to the prediction pattern is what was meant by the "conceptualization task" in the pattern matching process depicted in figure 3.4. Trochim (1989a, 1989b, 1989c) and Trochim and Linton (1986) have reviewed some alternative methods that might be used to accomplish this.

A third strategy for articulating the theoretical pattern can be called the *causal model method*. This approach is more deductive than the previous two and more directly links the theory with the theoretical expectation pattern. Here, stakeholders or theoreticians can be engaged in a process designed to articulate their causal models for a given context. The end of this process would be a causal map of the type shown in figures 3.1 through 3.3 or figure 3.5. From such causal diagrams, one can generate theoretical prediction patterns by assigning numerical values to the linkages and determining the hierarchies that result. A number of approaches for accomplishing such a task have been described in McClintock (1987), Trochim and Linton (1986), and Axelrod (1976), among others.

Assuming that we have a theory as implied in figure 3.5 and that we are able to generate a prediction pattern across the outcome variables, how might we carry out the study? Note that the model in figure 3.5 does not indicate the research design that might be used. We could construct a randomized experiment, a nonequivalent group design, or simply administer the treatment to a single group. The major trade-offs involve the degree to which each option helps rule out alternative explanations for the pattern of results. Regardless of the design that is chosen, one will ultimately need a numerical set of theoretical or expected outcome values across a set of outcome variables, and a second set of observed effect values (e.g., t, F, or β values) for the same variables. The key analysis involves the assessment of the degree to which these two patterns match or are associated with one another. Although more statistical work is needed to determine the best way to analyze such data, Trochim (1991) has suggested that a simple correlational test would be appropriate. Different correlational analyses might be used, depending on whether both patterns can be assumed as interval (i.e., Pearson product moment correlation) or ordinal (i.e., Spearman's Rho or the Tau correlation).

A PRELIMINARY EXAMPLE: A POST HOC PATTERN MATCHING STUDY OF THE EFFECTS OF A TRANSITIONAL EMPLOYMENT PROGRAM FOR YOUNG ADULTS WITH MENTAL ILLNESS

To illustrate some of the prior discussion about the role of pattern matching in theory-driven evaluation, a preliminary example of a post hoc pattern matching reanalysis study is offered. In many ways, this example is far from ideal. Although the outcome study was carried out several years ago, the theoretical pattern information was gathered from the program staff only recently. Furthermore, as will be seen later, there are several problems with the way in which the theoretical patterns were elicited that may lead to underestimating any potential pattern match. Nevertheless, the study illustrates the general pattern matching idea and shows some of the difficulties that a researcher is likely to encounter in trying to conduct such research. This example is presented only briefly here for expository reasons. A fuller description of the original evaluation can be found in J. A. Cook, M. Solomon and L. Mock (1989) and Cook, Solomon and J. A. Jonikas (1989). Trochim and Cook (1991) have described how pattern matching was used in this project in greater detail. The methodology is presented in two parts: a description of the original evaluation (development of the observed pattern) and a description of the subsequent involvement of the staff (the development of the theoretical pattern).

The Development of the Observed Pattern

The Young Adult Transitional Employment Program (hereafter abbreviated YA TEP) was created at the Thresholds Agency in Chicago, Illinois, which is a psychiatric rehabilitation agency with a special emphasis on comprehensive job placement and retention. The YA TEP is a youth program that is only one of many programs offered by the agency. The evaluation of the YA TEP was a relative one—the program consisted of six components that were *added to* the service delivery program that was already administered at the agency: (1) an "enhancing employability" course; (2) a community exploration program; (3) a visiting chefs program; (4) a vocational assessment battery; (5) a group placement for young adults only; and (6) a job club service (instruction and practice in job-hunting skills). The evaluation covered the project period from December 1, 1984, to November 30, 1988.

During the four-year study, 152 young adults entered the program. Of this group, 124 youths remained in the program longer than 90 days. Nearly three-quarters of the sample was male, over 90 percent were between 16 and 21 years of age, 73 percent had prior work experience, and all were considered mentally ill, with diagnoses including schizophrenia and affective disorders. The final sample consisted of 63 youths who were randomly assigned to receive either the new TEP treatment (treatment group N = 32) or the traditional agency service (control group N = 31) and who were measured at the beginning and end of the six-month training period and before job placement. Thus, the evaluation is essentially of the TEP training component, not its job placement effects.

All participants were measured on two occasions, once before beginning the program and once at the completion of the training phase and prior to placement. Although a larger battery of tests was actually administered, this pattern matching reanalysis limited the variables to the 86 items from six patient-rated psychosocial scales: the Self-Esteem Scale (Rosenberg, 1965; 10 items); the Coping Mastery Scale (Pearlin and Schooler, 1978; 7 items); the NAGI Index of Disability (Nagi, 1976; 13 items); the Stigma Scale (Neuhring, 1979; 16 items); the Zung Anxiety Scale (Zung, 1971; 20 items); and the Zung Depression Scale (Zung, 1965; 20 items). For each of the 86 items, a simple ANACOVA model was run to obtain univariate estimates of the relative effect of the treatment. There were two coding problems that needed to be addressed. First, on several scales, some of the items were deliberately reversed to avoid response sets. For instance, each item on the Rosenberg Self Esteem Scale is rated from 1 = Strongly Agree to 4 = Strongly Disagree. Item 2, "You feel that you have a number of good qualities," is a positive statement, whereas Item 3, "You are inclined to feel that you are a failure," is negative. For all negatively worded items, data were transformed as is commonly done for such reversals. The second problem came up in comparing across the six scales. On some scales, even when items were rescaled for reversals, the dominant direction for the scale was either positive or negative. Thus, all t-values for negatively oriented scales had to be reversed, so that all 86 items showed a "positive" treatment effect value (in this case, t-value) when there was a "clinically positive" movement. These 86 transformed t-values constitute the observed pattern for this pattern matching reanalysis.

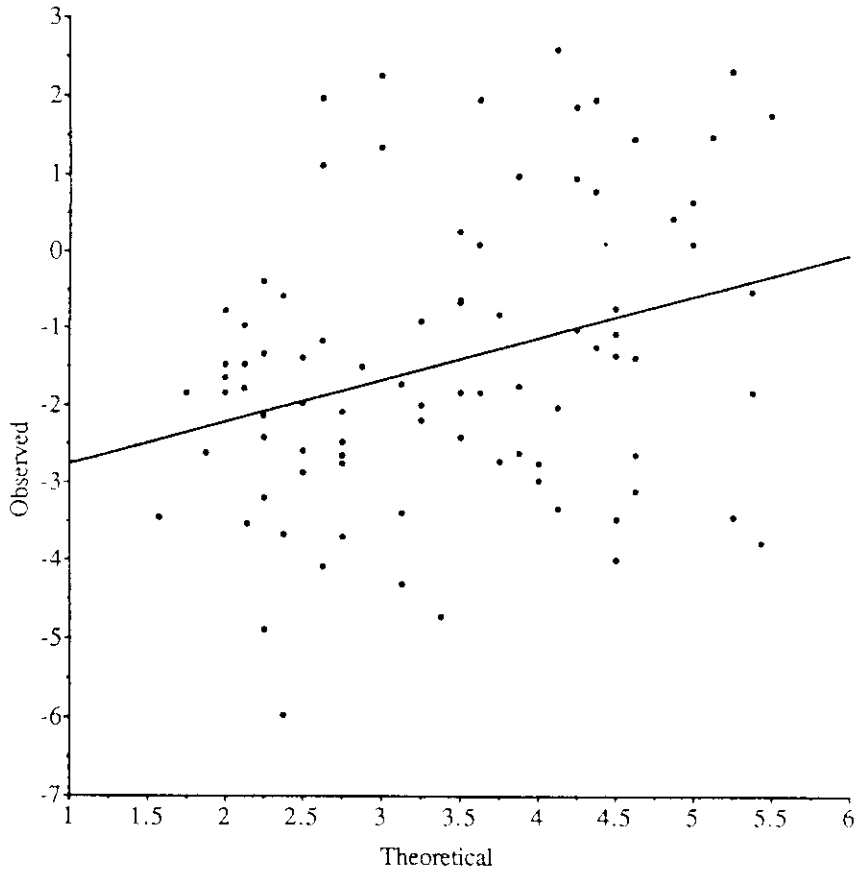
Development of the Theoretical Pattern

To develop the theoretical pattern, the direct rating method was used. Current staff members were asked to participate in a conceptualization process, the full details of which are reported in Trochim and Cook (1991) and will not be repeated here. Eight staff members participated in the data collection phase of relevance here. The key instruction of relevance here asked each staff member "to rate each of the 86 statements on a scale from 1 to 7, in terms of how much you think the Young Adult Program is likely to affect the attitude. If you think that the program will have a big effect on a specific attitude, you will give it a higher number; lower effects will be given low ratings." The ratings were then averaged across persons for each item to yield the theoretical expectation pattern.

Pattern Matching Results

The relationship between the theoretical and observed patterns is shown in figure 3.6. The theoretical values (average ratings of staff) are arrayed along the *x*-axis, while the observed values (item-level directionally-recoded t-values) are

Figure 3.6.
Bivariate Distribution for 86 Outcome Scale Items from Six Scales Showing Theoretical Pattern Estimates (Average Ratings of Expected Effect Magnitudes for Each Item from Program Staff) on the *X*-Axis and Observed Pattern Estimates (t-Values for Test of Differences between Treatment and Control Group Based on an ANACOVA Model) on the *Y*-Axis



along the *y*-axis. Higher observed values mean a greater gain in a clinically positive direction. The linear regression line is plotted through the data and shows a slight positive relationship between the theoretical and observed patterns. The overall Pearson product moment correlation for the 86 values is .30, which, though statistically significant at a .01 level, is hardly overwhelming.

There are several reasons to think that this pattern matching correlation may have been underestimated in this example. First, the rating directions that the

staff members followed in developing the theoretical expectation pattern were poorly worded and, based on subsequent conversations with the staff, probably misunderstood by at least some. The rating instructions do not correspond to the observed t-value estimates. The t-values constitute a relative comparison between the special TEP program and the usual program at that agency, but the rating instructions imply that an absolute treatment effect (that is, how effective TEP is relative to *no treatment at all*) should be rated. Second, the instructions for generating the theoretical expectation pattern ignored the directionality of effects. Under these instructions, staff members were asked only to rate the expected magnitude of the effect, but in reality any outcome could be either negatively or positively affected. The theoretical ratings do not distinguish these cases and consequently may reflect the expectations inaccurately. Finally, some of the staff reported that they thought they were being asked to rate how much change would be *desirable* on each item. For an item like Item 2 of the Self Esteem Scale --"You feel that you have a number of good qualities"—the staff gave a low rating of how much they thought the TEP program would affect it, even though in subsequent conversations, they acknowledged that they thought that item would be strongly affected. The misunderstanding arose because in the staff's view, if youths already felt that they "have a number of good qualities," then they would not want to change them at all! Therefore the staff gave a low rating for that item, even though they readily acknowledged that one would expect a positive treatment effect on it. Clearly, the potential for these kinds of misunderstandings in developing the theoretical pattern is important and likely to occur in almost any pattern matching situation. More work needs to be done on how to obtain clear theoretical patterns that are consistent with what the observed pattern is estimating.

One important implication concerns how a "positive" treatment effect is defined-- does a treatment that makes a young person express more "negative" self-assessments reflect a "positive" or "negative" outcome? Some would argue that lower postprogram self-esteem ratings should be taken as a negative treatment effect -- in other words, the person now has lower self-esteem than prior to treatment. But others would claim that youths who are more willing to admit to low self-esteem, coping mastery, and so on are actually taking the first hard step on the road to recovery from (or at least adaptation to) the disabling aspects of their illness. This view holds that a rational self-assessment ought to be negative and that youths who can recognize and admit their problems are showing the greatest "positive" effect, even though scale scores will show significant negative changes. This question is an important construct issue that needs to be addressed in any studies dealing with mental illness and related social or psychological measures. It is extremely important for pattern matching, but also needs greater recognition by all those engaging in theory-driven research.

This brief pattern matching example probably raises more questions than it answers. The fact that the theoretical pattern was solicited from staff long after the evaluation was completed immediately makes any pattern matching assess-

ment suspect. The real value of this example lies more in the methodological issues it raises than in the substantive answers it generates. Nevertheless, even under such dubious conditions, a low but significant pattern match was found.

The example illustrates a simple way to conduct outcome pattern matching that is consistent with current research practice and usually will not increase research costs substantially. The evaluator can generate estimates of treatment effect for a set of outcome variables using traditional methods, ideally doing item-level rather than scale-level analyses. Of course, it is preferable that these observed treatment effect estimates be based on randomized experimental designs, but they can be generated using quasi-experimental designs or even through simple pre-post measurement of a treatment group only. Staff members, administrators, social scientists, clients, and any other knowledgeable stakeholder groups can (preferably before the treatment) be involved in a conceptualization exercise for the purpose of generating their theoretical predictions of treatment effects for the same set of items. After the study, one can correlate the theoretical and observed patterns to see if there is evidence for a pattern match. We need many more examples of what happens in such studies before we will be able to make a more definitive judgment about the potential value of this type of pattern matching in theory-driven evaluations.

CONCLUSIONS

There is fairly widespread acceptance that the theory-driven critique of the more traditional "black box" experimental model is sensible. It is not yet clear that the theory-driven idea has moved beyond the level of a critique and has some reasonable alternative model to offer. It seems that the typical approach to formulating theory is based on the idea of the causal model. Yet causal modeling may not capture fully or well the potential of applying theory in program evaluation. Pattern matching is another approach to theory-driven evaluation. It *subsumes* the idea of causal modeling (because such models can be considered theoretical patterns) and suggests that we might *extend* that idea to include more fine-grained theoretical patterns and hunches. Proximal similarity is a useful, general theoretical principle that suggests that program outcomes should be similar for units—program variations, participants, or measures—that are more nearly alike. We can use this principle to generate more specific theoretical patterns that better reveal a unique "fingerprint" of the cause that we hope to find in the observations.

Another way to get at theoretical patterns is to involve stakeholders in the act of quantifying their expectations. The example provided here shows some of the difficulties involved in doing this that still need methodological attention. The example also indicates some special issues that arise in assessing the effects of rehabilitation services on severe mental illness. Since the process of rehabilitation may ideally need to involve recognition and acceptance of one's impairments, standardized scales and measures may show either *positive or negative* changes in self-perceptions, and these changes may be extremely difficult to interpret. Nevertheless, in many program evaluations, eliciting a theoretical expectation pattern will be relatively feasible and can be added on to the current analyses at little cost. With more studies of this nature, we will be in a better position to judge the merits of these types of pattern matching approaches.

NOTE

This research was funded in part by grants from the U.S. Department of Education, National Institute for Disability and Rehabilitation Research, the National Institute of Mental Health, Systems Development and Community Support (Grant #H133B00011) and the National Institute of Mental Health (Grant #MH46712 01A1).

REFERENCES

Axelrod, R. (1976). *Structure of decision: The cognitive maps of political elites*. Princeton, N.J.: Princeton University Press.

Bickman, L. (Ed.). (1987). *Using program theory in evaluation*. San Francisco: Jossey-Bass.

Bickman, L. (Ed.). (1990). *Advances in program theory*. San Francisco: Jossey-Bass.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W. Trochim. (Ed.), *Advances in quasi-experimental design and analysis* (pp. 67–78). San Francisco: Jossey-Bass.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Chen, H. T. (Ed.). (1989a). The theory-driven perspective (Special Issue). *Evaluation and Program Planning, 12*(4).

Chen, H. T. (1989b). The conceptual framework of the theory driven perspective. *Evaluation and Program Planning, 12*(4), 391–396.

Chen, H. T. (1990). *Theory-driven evaluations*. Newbury Park, Calif.: Sage.

Chen, H. T., & Rossi, P. H. (1980). The multi goal, theory driven approach to evaluation: A model linking basic and applied social science. *Social Forces, 59*, 106–122.

Chen, H. T., & Rossi, P. H. (1983). Evaluating with sense: The theory-driven approach. *Evaluation Review, 7*, 283–302.

Chen, H. T., & Rossi, P. H. (1987). The theory driven approach to validity. *Evaluation and Program Planning, 10*, 95–103.

Chen, H. T., & Rossi, P. H. (1989). Issues in the theory driven perspective. *Evaluation and Program Planning, 12*(4), 299–306.

Cook, J. A., Solomon, M. L., & Jonikas, J. A. (1989). Thresholds transitional employment program for mentally ill young adults: Final report to the U.S. Department of Education, Office of Special Education and Rehabilitative Services. Chicago: Thresholds Research Institute.

Cook, J. A., Solomon, M., & Mock, L. (1989). What happens after the first job placement: Vocational transitioning among severely emotionally disturbed and behavior disordered youth. *Programming for Adolescents with Behavior Disorders, 4*, 71–93.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Costner, H. L. (1989). The validity of conclusions in evaluation research: A further development of Chen and Rossi's theory-driven approach. *Evaluation and Program Planning, 12*(4), 345–366.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 4, 281–302.

Finney, J. W., & Moos, R. H. (1989). Theory and method in treatment evaluation. *Evaluation and Program Planning, 12*(4), 307–316.

Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. Cambridge, England: Cambridge University Press.

Kotz, S., Johnson, N. J., & Read, C. B. (1983). *Encyclopedia of the Statistical Sciences*. New York: Wiley, 462–466.

Lipsey, M. W., & Pollard, J. A. (1989). Driving toward theory in program evaluation: More models to choose from. *Evaluation and Program Planning, 12*(4), 317–328.

McClintock, C. (1987). Conceptual and action heuristics: Tools for the evaluator. In L. Bickman (Ed.), *Using program theory in evaluation*. San Francisco: Jossey-Bass.

Nagi, S. Z. (1976). An epidemiology of disability among adults in the United States. *Health and Society, 54*, 6–8.

Neuhring, E. M. (1979). Stigma and state hospital patients. *American Journal of Orthopsychiatry, 49*, 62–633.

Pearlin, L. I., & Schooler, C. (1978). The structure of coping. *Journal of Health and Social Behavior, 19*, 2–21.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton N.J.: Princeton University Press.

Runyan, W. K. (1980). A stage-state analysis of the life course. *Journal of Personality and Social Psychology, 6*, 951–962.

Scott, A. G., & Sechrest, L. (1989). Strength of theory and theory of strength. *Evaluation and Program Planning, 12*(4), 329–336.

Tinim, N. H. (1975). *Multivariate analysis with applications in education and psychology*. Monterey, Calif.: Brooks-Cole.

Trochim, W. (1985). Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review, 9*, 5, 575- 604.

Trochim, W. (1989a). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning, 12*(1).

Trochim, W. (Ed.). (1989b). Concept mapping for planning and evaluation. (Special Issue). *Evaluation and Program Planning, 12*(1).

Trochim, W. (1989c). Concept mapping: Soft science or hard art? *Evaluation and Program Planning, 12*(1).

Trochim, W. (1989d). Outcome pattern matching and program theory. *Evaluation and Program Planning, 12*(4), 355–366.

Trochim, W. (1991). Statistical analysis in outcome pattern matching. Unpublished manuscript, Cornell University.

Trochim, W., & Cook, J. A. (1991). A pattern matching assessment of the effects of a transitional employment program for mentally ill young adults. Unpublished manuscript, Cornell University.

Trochim, W., & Linton, R. (1986). Conceptualization for evaluation and planning. *Evaluation and Program Planning, 9*, 289–308.

Zung, W. W. K. (1965). A self-rating depression scale. *Archives of General Psychiatry, 12*, 63–70.

Zung, W. W. K. (1971). A rating instrument for anxiety disorders. *Psychomatics, 12*(6), 371–379.