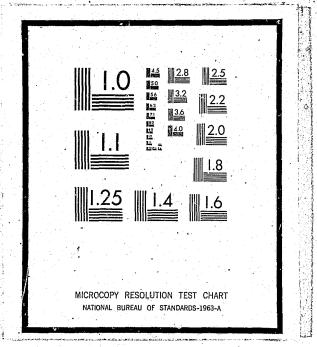# NCJRS

This microfiche was produced from documents received for inclusion in the NCJRS data base. Since NCJRS cannot exercise control over the physical condition of the documents submitted, the individual frame quality will vary. The resolution chart on this frame may be used to evaluate the document quality.

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Microfilming procedures used to create this fiche comply with the standards set forth in 41CFR 101-11.504

Points of view or opinions stated in this document are those of the author(s) and do not represent the official position or policies of the U.S. Department of Justice.

U.S. DEPARTMENT OF JUSTICE
LAW ENFORCEMENT ASSISTANCE ADMINISTRATION
NATIONAL CRIMINAL JUSTICE REFERENCE SERVICE
WASHINGTON, D.C. 20531

Date filmed 12/31/75

TABLE OF CONTENTS

PREFACE

Use of Program Evaluation is one of a series of Institute publications on mental health services research and development. The purpose of the series is to offer assistance to persons working toward continually increased effectiveness of delivering mental health contributions to people in need.

Reflected in all publications in the series is a three-phase process of services improvement through planning for creative change.

(1) Identification of problems and needs for change in services.

Use of Program Evaluation is aimed toward the improvement of formal approaches in front line facilities to help determine when change is —or is not—needed. The bibliography and abstracts provide an opportunity to review the current literature on program evaluation.

(2) Search and research to provide direction for effective change to solve problems and meet needs.

The publication, *Innovations and Current Conclusions,* issued several times each year, is to highlight innovative techniques. *Information Sources and How to Use Them* is offered as an aid to mental health workers seeking new knowledge through all relevant literature. A section of the *Manual on Research Utilization* has been addressed to those planning original research on innovative mental health services delivery techniques.
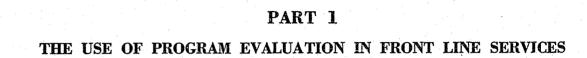
(3) Promotion of the diffusion and adoption of innovations through planned change.

Out of recognition that the dissemination of knowledge alone ushers little change, sections of the *Manual on Research Utilization* have been devoted to techniques of planned change, addressed to consultants and administrators/practitioners. For persons wishing to become more thoroughly familiar with the utilization of knowledge in planned change, *A Distillation of Principles on Research Utilization . . . Volume I* is offered. With the hope that it will foster continued investigations in refined techniques of change through knowledge utilization, *A Distillation of Principles on Research Utilization . . . Volume II—Bibliography with Annotations* has been issued as a part of this series.

The program evaluation bibliography and abstracts were prepared as part of the activities under Contract No. 42–69–82, National Institute of Mental Health, awarded to the Bureau of Applied Social Research, Columbia University. Very special thanks are due Carol Weiss, Karen Louis and Janet Weiss. The bibliography had not been required as a product of the contract. Ms. Weiss and her co-workers took the initiative in preparing the bibliography and annotations in response to the mounting number of requests for special material. Thanks are extended to Irma S. Lann, head of the NIMH Research Implementation Section, who served as project officer for the contract. But more than special thanks are owed her for originating the idea of the series and for serving as editor for all five of the publications.

Howard R. Davis, Ph.D.
CHIEF, MENTAL HEALTH SERVICES DEVELOPMENT BRANCH
DIVISION OF MENTAL HEALTH SERVICE PROGRAMS

PART 1

THE USE OF PROGRAM EVALUATION IN FRONT LINE SERVICES

THE USE OF PROGRAM EVALUATION IN FRONT LINE SERVICES

Evaluations carried out at the front line of action where mental health services are delivered might be distinguished by the term "hip pocket evaluations," unprestigious as that may sound. As excellence of the skilled craftsman's work results from continual measurement checks, so effectiveness of mental health services can grow with continual evaluations. A community mental health center may measure its own overall performance by methods of hip pocket evaluation. The contribution of a mental health program to community social change may be gauged. Or a clinician may use this concept of evaluation to identify the sorts of patients with whom he has greatest success.

Hip pocket evaluations are designed to fit within the operating budget of a service facility. They are usually self-evaluations. They may be conducted in conjunction with other staff or administrative duties. They are used by persons with varying degrees of experience with research methods. The purposes are to reinforce effective service and to signal the need for change in delivery techniques or policies.

The front line approach to evaluation is in contrast to the traditional examples of program evaluation generously supplied in leading references. (Williams and Ozarin, 1968; Bloom, 1970; Schulberg et al., 1969.) The program evaluation literature is largely devoted to major endeavors. Cited evaluations encompass a multitude of factors; commonly they call for full-time dedication of expert researchers. Heavy investment of funds is the rule. Proposals to NIMH for evaluation of mental health services entail yearly budgets averaging $75,000. It is small wonder that front line evaluations, as program management techniques, are somewhat less than routine. Practitioners and administrators seem to have developed a respect for the complexity of program evaluation that too often inhibits their direct involvement.

Neither does the front line approach encompass certain of the assessment techniques some people now place under the elastic rubric "program evaluation." For instance, utilization review, a procedure to monitor prompt and appropriate client care, meets a different sort of evaluation need and is not discussed in this brief guide. Patient records and biometric data processing systems meet still different needs. The use of social indicators as criterion measures of the impact of mental health programs is another technique hardly within the realm of practicability for front line evaluations. Systems evaluation of diverse consequences of program operations similarly goes beyond the more modest approach we are considering here.

Certainly the fact must be acknowledged that hip pocket evaluation is not a simple process to be applied without careful preparation and planning. Bad evaluation can be worse than no evaluation at all. But it is hoped that the suggestions, examples, and related references offered in this section will be of sound assistance to practitioners and administrators seriously interested in considering employment of program evaluation.

Benefits of frontline evaluation

There are at least three benefits which can accrue from front line evaluation efforts:

(1) Local program evaluation may be the key opening the way to continual refinement of services delivery. A 19th century German philosopher named Hermann Ebbinghaus asked students to draw quarter-inch lines. With hundreds of trials the lines only approximated one-quarter inch. But after having been given an ordinary hip pocket ruler the students achieved accuracy within three trials and were able to maintain it. In the same way program evaluation can improve performance.

The importance of program evaluation was stressed by Smith and Hobbs, who wrote in 1966: "The comprehensive community mental health center should devote an explicit portion of its budget to program evaluation. All centers should inculcate in their staff attention to and respect for research findings . . . only through explicit appraisal of program effects

can worthy approaches be retained and refined, and if ineffective, dropped."

(2) Program evaluation can provide needed reinforcement for practitioners. The nature of the human condition with which mental health is concerned often means changes are extremely subtle and gradual. Work loses much of its luster without feedback. "Evaluation can be fun!"

(3) As the impact of various service techniques becomes more commonly evaluated, determination of the respective values of alternative procedures will be made easier. If some standardization of local program evaluations can occur this advantage will, of course, be amplified. The pooled evaluation results from 30 facilities trying capitation financing, for example, would be considerably more cogent than the outcome of a single major demonstration or study.

Not without objections

Certainly criticisms of hip pocket evaluation have been leveled. One complaint is that self-evaluations lend themselves to insufficient objectivity. But probably only persons who would cheat at solitaire would cheat in evaluating their own programs! This is particularly so because the results of self-evaluation ordinarily are used by the evaluator himself rather than by a supraordinate group judging his program. Also, if the self-evaluation is conducted properly, an advisory body will be engaged.

Another problem is that local evaluations are considered by some to be a bit grubby because the results are not generalizable and therefore seldom publishable. But probably most practitioners and administrators place a higher value on eff' 're performance than on publication. Still, as local evaluations become more common and standardized, collaborative reports of cross-validated results with innovative techniques will represent rich contributions to the literature.

Twelve principles underlying soundness of hip pocket evaluations

Local service evaluations need be no less rigorous than major research undertakings. Observance of 12 principles can help insure high standards of measurement:

(1) Advisory groups should be utilized. Such bodies might be brought together only at the

planning and review stages of evaluation. Preferably the advisory group should include: representatives of the program's beneficiaries *and* supporters, such as higher organizational authority; key people from critical components of the rest of the system; and an appropriate consultant, if needed.

(2) Evaluations should be continual. They might be annual or timed with the adoption of a program innovation. The use of consistent criterion measures will allow the comparison of sequential evaluation results with base rates.

(3) A measuring device should be selected which will allow the reflection of better-than-expected performance. Most evaluation methods reflect only breaking even or losing, depriving staff of rewarding reinforcement.

(4) Parsimony is essential. One can become bogged down with excessive investments in recording and analyzing data that have no great relationship to decisions that can be made.

(5) Influence on the total system of one's agency should be considered rather than simply the straight line attainment of selected goals. The assessment of effects on other parts of the system may, of course, have to be carried out in a much more informal manner than the measurement of goal achievement.

(6) The evaluation should be conducted within the context of clear objective goal statements. One commonly comes upon service approaches that are followed simply because they seem to be the thing to do. If evaluation attempts did nothing more than sharpen fundamental objectives they would be worth their efforts.

(7) The evaluation should consider not only the attainment of the goal but what actions account for the attainment.

(8) Decisions and actions consequent to the evaluation should be planned at the outset. The critical question to ask in planning an evaluation is: "What might I do about the results?"

(9) To the extent possible, unobtrusive measures should be utilized. These include data that are already collected for other monitoring or reporting purposes. The more the data are standardized with those collected in similar facilities, the better the ultimate payoff of the evaluation. (Webb, E.J., Campbell, D.T., Schwartz, R.D., & Sechrest, L. *Unobtrusive measures: Nonreactive research in the social sciences.* Chicago: Rand McNally, 1966.)

(10) If special measuring instruments may be selected or devised, they should be checked for reliability (the likelihood that two persons rating the same event, even at different times, would agree) and validity (the similarity between what the instrument is supposed to measure and what it actually does measure). If comparison of outcomes is to be employed, "disinterested" raters should be obtained. They should be asked to rate examples from both comparison groups, but without knowledge of which group the instance represents.

(11) Classical experimental design—inappropriate if used slavishly—still stands as a guide to logical evaluation. Threats to the logical soundness of evaluations lurk continually—influences other than the actual service to be evaluated. These include: natural outside changes over time, such as seasonal influences; changes that happen to be occurring *within* persons studied for reasons that have nothing to do with the treatment; and biased selection of people or circumstances to be evaluated. Another example of hidden influences is called "statistical regression." It may take place if initial measurements are of extreme degrees. Virtually anything that is extreme at the outset will tend, on the average, to revert toward the less extreme.

The application of three safeguards will help ward off threats such as the above: (a) the use of either parallel activities of groups, or successive measurement of one group over repeated periods of time; (b) controlling the inputs to the comparison groups, by randomization when possible; and (c) checking for chance differences. Such simple statistical techniques as the "t test," "chi square," or "standard error of difference between percentages" can be called on to test for chance difference in a multitude of circumstances. Most psychologists can offer help with these, or the reader might wish to refer to other standard statistical references. (Thorndike, Robert L. and Elizabeth Hagen, *Measurement and evaluation in psychology and education*, 2nd edition. New York: John Wiley & Sons, 1961. Winer, B.J., *Statistical principles in experimental design.* New York: McGraw-Hill Book Co., Inc., 1962.)

(12) Distribution of the results of one's own evaluation is recommended even if that cannot be done through formal publication. Dissemination of evaluation results leads to wider interest

in one's program, anticipation of subsequent evaluation results, and, once one has put his own measure of effectiveness on the line, it provides an incentive for continual improvement.

An outline of suggested approaches

Because the range of uses of program evaluation is so vast, an outline of various approaches may provide a useful catalog at this point. References will be recommended for further reading where appropriate.

Method of Asking Clientele

An approach that lies above and beyond methodological rigor is that of informally asking a program's clientele to comment candidly on how they see the services offered. Beneficiaries are of primary concern, of course. But the supporters of services also constitute a most relevant clientele. The director of a State mental health research program in the Midwest learned it this way. He was faced with a problem: Earlier evaluation of the program had reflected its effectiveness in attracting and retaining quality staff who received grants for part-time research involvement. It also revealed that the number of publications had risen significantly. Despite such signs of payoff, the legislative appropriations for research had dwindled the previous two bienniums. At the annual research meeting the chairman of the Senate Appropriations Committee was invited to address the group on the subject "What I don't like about your research program." He had one complaint: the hard-wrought State tax money resulted in no manifest help to the patients served by the State mental health program. He was right. Diligent corrective efforts—subsequently reported to the Legislature—were accompanied by a 40-percent increase in the research appropriation the next session. Of course, a "causal" relationship was not necessarily established.

"Asking clientele" represents a sensible means to identify critical criteria. In the evaluation of the Institute's applied research grant program, the expressed views of terminating investigators led to awareness of the previously unthought-of needs to monitor continuity of staff contact and promptness of response to communications.

"The Behavior Modification" Method

One of the appeals of certain behavior modi-

fication techniques is the simple three-phase procedure. It is readily translatable to special types of program evaluation:

(1) Pinpoint one effort to be studied at a time.
(2) Devise a measurement of the outcome of the effort and repeat the measurement over intervals of time.
(3) Hold constant between measurements the techniques applied in the effort; employ new techniques until desired results are reflected by the measurements.

Formal Evaluation Methods

A logical classification of evaluation approaches has been offered by Tripodi et al: Monitoring, Research Techniques, and Cost-Analytic Techniques. (Tripodi, Tony; Fellin, Phillip, & Epstein, Irwin, *Social program evaluation*. Itasca, Illinois: Peacock, 1971.)

Monitoring

Accountability Audit—This is a common form of evaluation to meet the needs of boards, supporters, and State or national data banks. It includes maintenance of records on program expenditures, allocations, and the processing of beneficiaries. *General accounting* pertains to costs; *social accounting* pertains to such data as those on patient movements. Biometric reports are considered to fall under this definition. (Hill, John G., Cost analysis of social work service. Norman A. Polansky (Ed.), *Social work research*. Chicago: The University of Chicago Press, 1960.)

Administrative Audit—The determination of whether staff functions are being carried out according to predetermined standards is the function of this audit. (Schonfeld, H. K., Falk, I. S., Lavietes, P. H., Landwirth, I., & Krassnor, L. S., The development of standards for the audit and planning of medical care. *American Journal of Public Health*. November 1968, 88(11), 2097-2110.)

Time and Motion Studies—Though this allusion conjures up recollections of "Taylor troops" invading industry with their stopwatches and recording boards, time and motion studies can indeed lead to sharpening in the use of staff time in relation to their activities. (Elkin, Robert, *Analyzing time, costs, and operations in a voluntary children's institution and agency*. Washing-

ton, D.C.: Department of Health, Education, and Welfare, September 1965, 27-39.)

Research Techniques

Experiments—The practical experimental approach probably is the evaluation method of choice if one is to determine whether outcome is causally related to techniques, practices, and policies. These approaches are characterized by randomized inputs, alternative efforts or control and experimental groups, and the ascertainment of beyond-chance difference in results. Suchman offers a helpful description of several experimental paradigms. (Suchman, E. *Evaluative research: Principals and practice in public service and social action programs*. New York: Russell-Sage Foundation, 1967.)

Quasi-experimental designs—In front line service situations experimental control of circumstances is not always possible. (Actually, such may be possible more often than it seems.) Quasi-experimental designs take advantage of what opportunities there are to control against threats to logical soundness.

Time series—Successive measures of output are made, commonly with different groups on the assumption that the flow of clients into the program remains unchanged over time.

Multiple time series—This variation refers to the fact that *successive* measurements are obtained, with statistical comparisons between and among them.

Nonequivalent controls—No effort is made to randomize or match the clients into two groups. However, the reasonably most similar hospital ward, for example, may be selected as a comparison group.

Patchwork designs—One controls whatever is necessary and possible to control. For instance, if a practitioner feels that age is really the only factor that would account for differences in the response of patients to a special technique, he would try to have both comparison groups matched in terms of age; but he would make no effort to control other variables.

A good rule of thumb when using quasi-experimental designs is always to ponder other plausible explanations, discounting them by logical assessment if possible. An excellent discussion of experimental designs can be found in Campbell & Stanley. (Campbell, Donald T., & Stanley, Julian C. *Experimental and quasi-ex-*

perimental designs for research. Chicago: Rand McNally, 1969.)

Surveys

One usefulness of surveys is the assessment of a mental health program's impact on a total community. Attitudes, opinions, and reported changes in behavior are common factors measured. Of course, survey data can be subjected to classical experimental designs. (Glock, Charles Y. (Ed.) : *Survey research in the social sciences*. New York: Russell-Sage Foundation, 1967.)

Case Studies

Detailed descriptions are prepared for groups of cases. The descriptions are subsequently reviewed for clusters of facts which begin to emerge as meaningful patterns. In order to learn what the differences were between research projects which had high payoff and those which had low payoff, one of the NIMH research grant programs supported a case study of a number of terminated projects. Samples of projects of both extremes, matched on topic, investment of funds, and duration, were visited. The analysis of the case descriptions based upon extensive interviews yielded 15 factors that differentiated high payoff from low payoff projects. The results were utilized in subsequent proposal reviews.

Cost Analytic Techniques

Cost accounting—This approach relates program costs to output costs. For instance, the dollars invested per patient released from the mental hospital is sometimes used to compare various hospitals within a program, or one hospital with itself over a period of time. In other instances the cost accounting results will read something like this: "X dollars were used for Y amount of man-hours to reach goals for Z number of patients." Obviously, cost accounting in program evaluation depends upon clear program objectives and categories of unit evaluation.

Cost-benefit analysis—The relative effectiveness of *alternative* programs, strategy, etc., are measured. This method differs from cost accounting in that alternative approaches are compared. The primary concern may be determination of the resources required to meet a specific goal. This approach expresses goal at-

tainment in dollars—the economic productivity of patients, for example. (Levine, Abraham S., Cost-benefit analysis in social welfare: An exploration of possible applications. *Welfare in review*. February 1966 (4) 2, 1-11.)

Cost-outcome analysis—Program goals are related, not in dollars but in terms of other specified criteria. This method, given its name by Tripodi, et al., may hold the output constant. For example, the output may be specified as the "release of patients within 21 days." Then various inputs would be experimented with to see which one could achieve the desired results with the least investment.

Operations research—"OR" pertains to alternate ways of conducting and coordinating program activities within an agency. It employs systems techniques, mathematics, and computer science. It can be of considerable advantage in evaluations toward more appropriate assignment of staff, the scheduling of patients, choosing among program options, etc. The Halpert et al. monograph on Operations Research in Mental Health is a fruitful resource. (Halpert, Harold P., Horvath, William J., & Young, John P. *An administrators handbook on the application of operations research to the management of mental health systems*. Washington, D.C.: Public Health Service Publication No. 2110.)

Models for approaching program evaluation: two examples

Two program evaluation models which currently are receiving wide attention are *Key Factor Analysis* and *Goal Attainment Scaling*. Both will be outlined to illustrate their features.

Key Factor Analysis (no relationship to the statistical technique of factor analysis) was developed by Irwin M. Jarett, Ph.D., Chairman, Departments of Accounting and Finance, Southern Illinois University. Technically, it may be described as an application of general systems theory to organization and to management. A special asset of Key Factor Analysis is that not only is it a method of program evaluation offered but a systematic route toward program planning is inherent.

Planning for a Key Factor Analysis is carried out under an assumption that the organization has no commitment to an already established program. Hence the expression, "program-free

planning." The planning group ideally should consist not only of staff but of beneficiaries and supporters of the organization.

Planning takes place over an eight-step process:

Step 1. *Purpose*—The group first considers what human needs and what populations the organization exists to serve. (Planning may start within a subcomponent of a larger organization. Conceivably, a hospital ward could be considered an organization.)

Example: Let the "organization" be the consultation staff of a community mental health center. The *Purpose* might be "to provide consultation to community agencies toward the mental well-being of all citizens served by those agencies."

Step 2. *Objective*—Stated in output terms, this is an individual statement of the needs included in *Purpose*.

Example: "Ten percent increase in consultation to community agencies."

Step 3. *Objective groupings*—In a natural planning situation the persons involved would submit as many objectives related to achievement of the purpose as came to mind. These then would be clustered if some seemed to be saying the same thing. If 25 objectives were suggested, they might be expected to condense down to five objective groupings, which would then be treated as specific objectives.

Step 4. *Key factor*—A success or failure criterion for an objective grouping.

Example: *Amount* of consultation.

Step 5. *Key indicator*—A specific measure which constitutes support for the definition of the key factor noted above. It represents in program evaluation terms the "criterion measure."

Example: Number of consultation events during a given period of time.

Step 6. *Goals*—Time-limited, organization-oriented statements of intended progress toward the specific objectives.

Example: Orientation sessions with community agencies and demonstrations of effective consultation services.

Step 7. *Program*—The collection of resources for the express purpose of achieving the goal.

Example: Provision of competent consultants with adequate time allocation for serving community institutions.

Step 8. *Management Information System*—

This is a repository for all information listed above. This system provides for the analysis of results, as yielded by the key indicators. Data on these specific measures may, of course, be subjected to any experimental or quasi-experimental method for interpretation. This step might be considered the actual locus of program evaluation machinery.

Excellent discussions of Key Factor Analysis may be found in: Western Interstate Commission for Higher Education. *Systems approach to program evaluation in mental health;* Boulder, Colo.: WICHE, 1970.

Goal Attainment Scaling was originated by Thomas Kiresuk, Ph.D., Chief Clinical Psychologist and Director of Research and Program Evaluation, Hennepin County Mental Health Center, Minneapolis, Minn. "GAS" provides an estimate of whether or not the goal was actually reached which someone thought would be reached. It is a particularly versatile model, lending itself to an almost limitless variety of goals related to clinical, program services, or administrative activities. Properly used, it can satisfy most standards and principles of rigorous program evaluation. Goal Attainment Scaling offers two advantages rarely found in other models: (1) One can compare the attainment of one goal with that of any other goal, even though different criteria of attainment have been used. (2) The attainment of better-than-expected success can be reflected. This is in contrast to the more customary break-even-or-lose outcome. Consequently, it allows an opportunity for reward for success.

Use of the model occurs in three phases:

Phase 1. The user prepares an *objective* statement of what he thinks he (or the program activity being evaluated) will accomplish at a given point in time with regard to a selected task. Then he briefly describes what the situation should be at the same point of time if all breaks are in favor of pursuing this particular goal. The same thing is done assuming that all breaks are against him. Between the most likely achievement and the very best outcome that could be described the user tries to objectively state a midpoint outcome. The same thing is done in the direction of underachieving the goal. This results in a five-point scale: −2, −1, 0, +1, +2. The statements are placed on a grid as illustrated

below. If more than one goal is of concern in the process each separate goal may be given a weight. The weight is determined in accordance with the importance of that particular goal relative to any other goals on the grid.

Table I
KIRESUK GOAL ATTAINMENT FOLLOWUP GUIDE

Followup Data: (12 months) Alpha Mental Health Center.

GOAL ATTAINMENT FOLLOWUP GUIDE

Scale attainment levels	Scale headings and weights	
	Scale 1: Consultation contacts ($W_1 = 5$)	Scale 2: Requests by multiple agencies ($W_1 = 7$)
a. Most favorable outcome thought likely. (+2)	*+ 30% <u>X</u>	School, social service, police, physicians.
b. More than expected success. (+1)	+ 20%	Three of the above.
c. Expected level of success. (0)	+ 10%	Two of the above.
d. Less than expected success. (−1)	0%	One of the above. <u>X</u>
e. Most unfavorable outcome thought likely. (−2)	− 10%	No regular requests, only preferred consultation.

* % change

Phase 2. At a predetermined point in time, the actual state of affairs with regard to each goal is measured or observed. An indication of the appropriate rating is made on the grid. <u>X</u>

Phase 3. In order to compare and interpret goal attainments, the "standard score" is calculated.* The "S" score is a statistical tool which allows one to compare two scores on the

---

* *The formula for calculation is:*

$$\text{"S" SCORE} = 50 + \frac{10 \sum\limits_{i=1}^{n} w_i x_i}{\sqrt{1-P \sum\limits_{i=1}^{n} w^2_i + P \left(\sum\limits_{i=1}^{n} w\right)^2}}$$

---

basis of how far they are from "average." In this case, one "pretends" that hundreds of tries have been made at achieving this particular goal. The average has, presumably, turned out to be at the "zero" level on the grid. (Recently Kiresuk and co-workers checked the soundness of this assumption in reality. They found that it is indeed valid to so assume, *if* the goal prediction statements have been made with good judgments.) Here are the steps for figuring the standard scores.

Step One. For each scale, multiply the rating times the weight.

In our example: $(+2) \times (5) = 10$; $(-1) \times (7) = -7$.

Step Two. Add up the answers for as many scales as have been used on the grid.

In our example: $(10) + (-7) = 3$.

Step Three. Multiply result of Step Two by 10 (a statistical maneuver).

Example: $10 \times 3 = 30$.

Step Four: Square each weight, and add up.

Example: $5^2 = 25$; $7^2 = 49$; $25 + 49 = 74$.

Step Five: Multiply Step Four by .7 (a statistical maneuver).

Example: $74 \times .7 = 51.8$

Step Six: Add up the weights and square that sum.

Example: $5 + 7 = 12$; $12^2 = 144$

Step Seven: Multiply results of Step Six by .3 (a statistical maneuver).

Example: $144 \times .3 = 43.2$

Step Eight: Add results of Steps Five and Seven.

Example: $51.8 + 43.2 = 95.0$

Step Nine: Extract the square root of result of Step Eight.

Example: $\sqrt{95} = 9.74$

Step Ten: Divide result of Step Nine into result of Step Three.

Example: $\dfrac{30}{9.74} = 3.08$

Step Eleven: Add 50 to result of Step Ten (a statistical maneuver).

Example: $50 + 3.08 = 53.08$

The "S" score turns out to be 53.08.

To interpret: An "S" score of 50 would have meant attainment on the whole of just what was predicted.

An "S" score of 60 (or 40) would occur by chance alone (without its being due to special

performance) only once in 68 tries; an "S" score of 70 (or 30) would occur by chance alone once in 99 tries. (An "S" score of 60 represents a +1 on the grid; an "S" score of 70 represents a +2.)

Scale 1, *Consultation Contacts*, also could be compared with Scale 2, *Requests by Multiple Agencies*. In actual calculation Scale 1 results in an "S" score of 73.3, compared with Scale 2 "S" score of 39.9, a finding reflecting that intensified efforts to push consultation failed to evoke requests for such service.

Results of GAS can, as was true for Key Factor Analysis measurements, be subjected to experimental or quasi-experimental designs.

It has been suggested that the advantages of both KFE and GAS might be marshalled by treating *Key Indicators* as 0-level goals in the GAS grid. (Kiresuk, T.J., & Sherman, R.E. Goal Attainment Scaling: A general method of evaluating comprehensive community mental health programs. *Community Mental Health Journal*, 1968, 4, 443–453. (Recent bulletins are available from Mrs. Susan Salasin, Assistant Director, PEP, Minneapolis Medical Research Foundation, McGill Building, Minneapolis, Minn.)

Other special models

Continuous Monitoring of Outcome

An approach to the evaluation of community mental health center programs which might be considered a true "model" is being developed by Dr. James A. Ciarlo, Community Mental Health Center, Denver General Hospital. The system continually assesses, in terms of the outcome of treated clients, the benefits in relation to program services offered during a specified period of time. The system also addresses indirect client services from the standpoint of impact on other care giving agencies. In its most elaborate form, the Ciarlo system probably extends beyond what we have been considering as resources available for front line evaluations. However, less ambitious implementation is perhaps possible using staff and processing facilities already available in most centers. Even though the system is in the early stages of being tested, it is attracting an extraordinary amount of nationwide interest.

Dr. Ciarlo has kindly consented to our in-

cluding the following outline, developed for discussion purposes:

I. General Characteristics of the Evaluation Systems as of January 1972.

A. *Direct* client services:

1. Evaluation is focused upon program *outcome*, in terms of the mental health of *treated clients*. The outcome assessed include:

   a. How much *psychological discomfort* is the client now experiencing?
   b. How *interpersonally isolated* is the client now?
   c. How *productive* (in a job, in housework, in school) is the client?
   d. To what degree is the client now *abusing alcohol* or *drugs*?
   e. Is the client in *trouble with the law* (arrests for drunkenness, drug possession, etc.)?
   f. To what degree is the client now *dependent upon public services* to maintain him (psychiatric, custodial, welfare, etc.)?
   g. How *satisfied* is the client with the services he has received?

2. Evaluation does *not* focus on CMHC *process* variables (such as number of hours of care provided, etc.), but it does focus also on three process-related system characteristics:

   a. Is care *accessible* to all persons in the catchment area?
   b. Are clients moved effectively between treatment modalities, as the need arises (*continuity of care*)?
   c. Does the service system *minimize* the flow of patients to *long-term care*, away from their own home, job, etc.?

3. Evaluation does *not* focus on incidence rates, prevalence rates, or other "social indicator" rates for either the catchment area or the larger community. Use of such rates as outcome indicators is believed to obscure, rather than illuminate the true effectiveness of CMHC services (particularly specific treatment technique effectiveness).

4. The outcome measures used can be applied across *all* treatment modalities (inpatient, day care, emergency, etc.), so

that the relative effectiveness of each, or combinations, can be determined.

5. If a CMHC *program* is defined in terms of client characteristics (e.g. alcoholics, addicts, chronic psychotics, etc.), or in any manner which identifies clients served by a program, outcome levels achieved by a program can be determined at any time. Also, the largest and smallest contributors to that outcome level can be identified for appropriate action by management.

6. A principal evaluation focus is on the *change in program outcome levels* over time, so that the CMHC knows whether its effectiveness is increasing, decreasing, or remaining unchanged, and can take appropriate action (remedial or "experimental"). This is referred to as *continuous monitoring of outcome*.

7. Cost of services can be related to program outcome levels, to determine (a) whether the benefits-to-cost ratio is increasing, decreasing, or staying constant; and (b) to determine the more or less economical service patterns which may produce a given outcome level.

B. *Indirect* client services:

1. Evaluation focuses upon the *impact on other caregivers* (satisfaction level, attitudes, behavior). It does not focus upon the impact on the caregiver's *clients*, even though that impact is the crucial one. However, the difficulties in studying persons *outside* the CMHC system are formidable, and preclude continuous assessment of the actual psychological status of such people. Caregiver "feedback" and functioning was selected as a more feasible evaluation focus.

2. If great care is used in selecting and refining certain incidence or "social indicator" rates (e.g. new school dropouts), changes in such rates *may* be useful as indicators of the *prevention* effectiveness of a program aimed at altering that rate.

II. System Implementation Requirements

A. Management support:

1. *Commitment* to implementing the system and providing the necessary resources. About 5 percent of available staff time

is believed to be a minimum commitment of resources.

2. Administrative *facilitation* of data-gathering, record-keeping.

3. Willingness to use evaluation results by *acting* on them. The usual "program justification" motivation is inadequate and will not support true evaluation.

B. Personnel:

1. At least one research-trained staff member to supervise data-collection and analysis procedures.

2. At least one person able to locate clients and ex-clients in the community, and to conduct a standard follow-up interview.

C. Records system:

1. A *reliable* record-keeping system which records *by client* (a) demographic characteristics of clients, (b) the problems (including diagnoses) of clients, (c) the services rendered to each client (including types, dates, dosages, serving clinician, unit, etc.), and (d) disposition and plans for further care. These are also the minimum requirements for a good *clinical* records system, and a single system may serve both clinical and evaluative purposes.

2. Capability of retrieving any variable mentioned above (sex, diagnosis, etc.) in relationship to any other variable. This capability is essential for drawing comparable groups for later follow-up, statistically controlling a variable in an analysis, etc. A punch-card system (McBee, IBM) is probably necessary for small CMHCs; computer service is probably essential in large ones (those seeing 5,000 or more patients annually.

Dynamic Evaluation

For evaluation of a program as a system in the process of continual change see Parsell, Alfred T. *Dynamic Evaluation: The systems approach to action-research*. Professional paper No. SP–2423;, Santa Monica, California: System Development Corporation, 1966.

Differential Evaluation

For evaluation of a program yet to be launched, assessing stages of initiation, establishing contacts with clients, and implementa-

tion, see Tripodi, Tony, Fellin, Phillip, & Epstein, Irwin, *Social Program evaluation;* Itasca, Illinois; Peacock, 1971.

Program Effectiveness Evaluation

For evaluation of progress toward an ultimate objective, when sequential subobjectives are measurable, see Deniston, O. L., Rosenstock, I. M., & Getting, V. A. Evaluation of Program effectiveness. *Public Health Reports,* 1968, 4 (83) 323-335.

A view on new models

Marcia Guttentag presents some refreshing thoughts with which everyone engaged in program evaluation should be aware. (Guttentag, Marcia, "Models and Methods in Evaluation Research," *Journal for the Theory of Social Behavior,* Vol. 1, No. 1, (1971), 75-95.) She appropriately assails our tired tendency to resort to the classical experimental design in program evaluation. Alternative models are suggested. One is the *Legal Model,* adopting to a degree the rules which govern the presentation and evaluation of evidence from the legal system. A greater amount of relevant data can be considered than in the classical experimental model. Another model is the *Decision Theoretic Approach* which gives regard to personal probabilities rather than frequentistic probabilities. Employing Bayesian statistics, one can stop data collection at any time, analyze data already collected, make decisions, revise the program, and continue on. Guttentag goes on to describe methods to encompass *context* in program evaluation: *Eco-behavioral units* and *Social Area Analysis.* The first considers person-place interactions—unquestionably an important determinant of program impact. That

interaction rarely is formally considered in program evaluation. The second method is one of the most promising approaches to the general field of social indicators. It has been brought to a level of refinement recently, largely through the work of Elmer Streuning.

Guttentag's assertions are likely to constitute a prophecy. Willy-nilly the reluctance of many evaluators to use social indicators as criteria increases. California's new Lanterman Law requires it in the evaluation of State-supported mental health programs. On the other hand it may be premature for the bell to toll for the classical experimental approach to program evaluation in front line operations. The models Guttentag describes are really not yet available in a widely utilizable form. This refers not only to the fact that they need to be shaped and refined by research in broader circumstances, but that people in the field of mental health need more time to grow comfortable with the concepts behind them. The classical approach fits more easily with our customary, even simpler, ways of thinking. That virtue—and, of course, it is also a liability—very likely renders the results of evaluations easier to utilize by most decision makers right now. So even though close attention should be paid to the more appropriate "second generation" models and methods being explored, it seems that for the sake of wide implementation and utilization of front line evaluation, we will have to use what we have and what we know. The important point to keep in mind is that the familiar classical methods are indeed limiting, and that more appropriate, flexible, and efficient methods should be assimilated when that is feasible.

H. R. Davis

PART II

BIBLIOGRAPHY ON EVALUATION RESEARCH

Part II, *Bibliography,* and Part III, *Abstracts,* were prepared by Carol Weiss, Karen Louis, and Janet Weiss.

BIBLIOGRAPHY ON EVALUATION RESEARCH

1. Conceptual and Methodological Issues

*Alkin, Marvin C., "Evaluation Theory Development," *Evaluation Comment*, vol. 2, No. 1 (1969), 2–7.

American City Corporation, *Making Evaluation Research Useful: A Summary of Proceedings of a Workshop*, Columbia, Md: Urban Life Center, 1971.

*American Institutes for Research, *Evaluative Research Strategies and Methods*, Pittsburgh: American Institutes for Research, 1970.

*Andrew, Gwen, "Some Observations on Management Problems in Applied Social Research," *The American Sociologist*, vol. 2, No. 2 (1967), 84–89, 92.

Archibald, K. A., "Three Views of the Expert's Role and the Clinical Approach," *Policy Sciences*, vol. 1 (1970), 73–86.

———, "Alternative Orientations to Social Science Utilization," *Social Science Information*, vol. 9, No. 2 (1970), 7–34.

*Aronson, Sidney H. and Clarence C. Sherwood, "Researcher versus Practitioner: Problems in Social Action Research," *Social Work*, vol. 12, No. 4 (1967), 89–96.

Baker, Robert L., "Curriculum Evaluation," *Review of Educational Research*, vol. 39, No. 3 (1969), 339–358.

*Barton, Allen H., *Studying the Effects of a College Education*. New Haven: The Edward H. Hazen Foundation, 1959.

*Bateman, Worth, "Assessing Program Effectiveness: A Rating System for Identifying Relative Program Success," *Welfare in Review*, vol. 6, No. 1 (1968), 1–10.

*Belshaw, Cyril S., "Evaluation of Technical Assistance as a Contribution to Development," *International Development Review*, vol. 8 (June, 1966), 2–23.

*Benedict, Barbara A., Paula H. Calder, Daniel M. Callahan, Harvey Hornstein, and Matthew B. Miles, "The Clinical-Experimental Approach to Assessing Organizational Change Efforts," *Journal of Applied Behavioral Science*, vol. 3, No. 3 (1967), 347–380.

*Bennis, Warren, "Theory and Method in Applying Behavioral Science to Planned Organizational Change," *Journal of Applied Behavioral Science*, vol. 1, No. 4 (1965), 337–360.

Berlak, Harold, "Values, Goals, Public Policy and Evaluation," *Review of Educational Research*, vol. 40, No. 2 (1970), 261–278.

Bigman, Stanley K., "Evaluating the Effectiveness of Religious Programs," *Review of Religious Research*, vol. 2, No. 3 (1961), 97–121.

*Blenkner, Margaret, "Obstacles to Evaluative Research in Casework," parts 1 and 2, *Social Casework*, vol. 31, Nos. 2 and 3 (1950), 54–60, 97–105.

*Bloom, Bernard L., "Mental Health Program Evaluation," In: S. Golann and C. Eisdorfer (Eds.), *Community psychology and mental health*, New York: Appleton-Century Crofts, 1970.

Blum, Hendrik L. and Alvin R. Leonard, "Evaluation Research and Demonstration," in *Public Administration: A Public Health Viewpoint*, pp. 286–322. New York: The Macmillan Co., 1963.

*Borgatta, Edgar, "Research: Pure and Applied," *Group Psychotherapy*, vol. 8, No. 3 (1955), 263–277.

*———, "Research Problems in Evaluation of Health Service Demonstrations," *Milbank Memorial Fund Quarterly*, vol. 44, No. 4 (1966), part 2, 182–199.

Brim, Orville, G., Jr., "Evaluating the Effects of Parent Education," *Journal of Marriage and Family Living*, vol. 19 (February, 1957), 54–60.

*Brooks, Michael P., "The Community Action Program as a Setting for Applied Research," *Journal of Social Issues*, vol. 21, No. 1 (1965), 29–40.

*Brunner, Edmund deS., "Evaluation Research in Adult Education," *International Review of Community Development*, No. 17–18 (1967), 97–102.

*Bynder, Herbert, "Sociology in a Hospital: A Case Study in Frustration," in *Sociology in Action*, edited by Arthur B. Shostak, pp. 61–70. Homewood, Illinois: Dorsey Press, 1966.

*Caldwell, Michael S., "An Approach to the Assessment of Educational Planning," *Educational Technology*, vol. 8, No. 19 (1968), 5–12.

Campbell, Donald T., "Administrative Experimentation, Institutional Records, and Nonreactive Measures," in *Improving Experimental Design and Statistical Analysis*, edited by J. C. Stanley, pp 257–291. Chicago: Rand McNally, 1967.

*———, "Reforms as Experiments," *American Psychologist*, vol. 24, No. 4 (1969), 409–429.

*———, "Considering the Case Against Experimental Evaluations of Social Innovations," *Administrative Science Quarterly*, vol. 15, No. 1 (1970), 110–113.

*Caro, Francis G., "Approaches to Evaluative Research: A Review," *Human Organization*, vol. 28, No. 2 (1969), 87–99.

——— (Ed.), *Readings on Evaluative Research*, New York: Russell Sage Foundation, 1971.

Cherns, A., "The Use of the Social Sciences," *Human Relations*, vol. 21, No. 4 (1968), 313–325.

Cohen, David K., "Policies and Research: Evaluation of Program Effectiveness," *Public Health Reports*, vol. 40, No. 2 (1970), 213–238.

*Colvin, C. R., "A Reading Program that Failed—or Did It?" *Journal of Reading*, vol. 12, No. 2 (1968), 142–146.

*Community Council of Greater New York, Research Department, *Issues in Community Action Research*, Report of the Spring Research Forum on Evaluation

Efforts in Three New York City Community Action Programs, 1967.

*Cronbach, Lee, "Evaluation for Course Improvement," *Teachers College Record*, vol. 64, No. 8 (1963), 672–683. Also reprinted in *Readings in Measurement and Evaluation*, edited by Norman Gronlund, pp. 37–52. New York: The Macmillan Company, 1968.

Daily, Edwin F. and Mildred A. Morehead, "A Method of Evaluating and Improving the Quality of Medical Care," *American Journal of Public Health*, vol. 46, No. 7 (1956), 848–854.

*Davis, James A., "Great Books and Small Groups: An Informal History of a National Survey," in *Sociologists at Work: Essays on the Craft of Social Research*, edited by Phillip E. Hammond, pp. 212–234. New York: Basic Books, 1964.

Deniston, O. L., I. M. Rosenstock, and V. A. Getting, "Evaluation of Program Effectiveness," *Public Health Reports*, vol. 83, No. 4 (1968), 323–335.

———, I. M. Rosenstock, W. Welch, and V. A. Getting, "Evaluation of Progress Efficiency," *Public Health Reports*, vol. 83, No. 7 (1968), 803–610.

———, I. M. Rosenstock, "Evaluating Health Programs," *Public Health Reports*, vol. 85, No. 9, (1970), 835–840.

*Dexter, Lewis A., "Impressions About Utility and Wastefulness in Applied Social Science Studies," *American Behavioral Scientist*, vol. 9, No. 6 (1966), 9–10.

*Donabedian, Avedis, "Evaluating the Quality of Medical Care," *Milbank Memorial Fund Quarterly*, vol. 44, No. 3 (1966), part 2, 166–203.

*Dorfman, Robert, "Introduction," in *Measuring Benefits of Government Investments*, pp. 1–11. Washington, D.C.: Brookings Institution, 1965.

*Downs, Anthony, "Some Thoughts on Giving People Economic Advice," *American Behavioral Scientist*, vol. 9, No. 1 (1965), 30–32.

*Dressel, Paul L. (ed.), *Evaluation in Higher Education*. Boston: Houghton Mifflin Company, 1961.

*Drew, Elizabeth B., "HEW Grapples with PPBS," *The Public Interest*, No. 8 (Summer 1967), 9–29.

*Dyer, Henry S., "The Pennsylvania Plan: Evaluating the Quality of Educational Programs," *Science Education*, vol. 50, No. 3 (1966), 242–248.

*Eaton, Joseph W., "Symbolic and Substantive Evaluation Research," *Administrative Science Quarterly*, vol. 6, No. 4 (1962), 421–442.

Educational Evaluation: New Roles, New Means, The 68th Yearbook of the National Society for the Study of Education, edited by Ralph W. Tyler. Chicago: National Society for the Study of Education, 1969.

*Educational Testing Service, *On Evaluating Title I Programs*. Princeton, N.J.: ETS, 1966.

Eidell, Terry L. and Joanne M. Kitchel, *Knowledge Production and Utilization in Educational Administration*. Eugene: Center for the Advanced Study of Educational Administration, University of Oregon, 1968.

*Elinson, Jack, "Effectiveness of Social Action Programs in Health and Welfare," in *Assessing the Effectiveness of Child Health Services*, Report of the

Fifty-Sixth Ross Conference on Pediatric Research, pp. 77–78. Columbus, Ohio: Ross Laboratories, 1967.

Etzioni, Amitai, "Two Approaches to Organizational Analysis: A Critique and a Suggestion," *Administrative Science Quarterly*, vol. 5, No. 2 (1960), 257–278.

*——— and Edward W. Lehman, "Some Dangers in 'Valid' Social Measurement," *The Annals of the American Academy of Political and Social Science*, vol. 373 (September 1967), 1–15.

*"Evaluating Educational Programs: A Symposium," *Urban Review*, vol. 3, No. 4 (1969), 4–22.

*Evans, John W., "Evaluating Social Action Programs," *Social Science Quarterly*, vol. 50, No. 3 (1969), 568–581.

*Fairweather, George W., *Methods for Experimental Social Innovation*, pp. 24–36. New York: John Wiley and Sons, 1967.

*Fellin, Phillip, Tony Tripodi, and Henry J. Meyer (eds.), *Exemplars of Social Research*. Itasca, Illinois: F. E. Peacock Publishers, 1969.

*Ferman, Louis A., "Some Perspectives on Evaluating Social Welfare Programs," *Annals of the American Academy of Political and Social Science*, vol. 385 (September 1969), 143–156.

First National Conference on Evaluation in Public Health. Ann Arbor: University of Michigan, School of Public Health, Continued Education Series, No. 89, 1960.

*Flanagan, John C., "Evaluating Educational Outcomes," *Science Education*, vol. 50, No. 3 (1966), 248–251.

———, "Project Talent: The First National Census of Aptitudes and Abilities," in *Readings in Measurement and Evaluation*, edited by Norman Gronlund, pp. 413–421. New York: The Macmillan Co., 1968.

*Fleck, Andrew C., Jr., "Evaluation as a Logical Process," *Canadian Journal of Public Health*, vol. 52, No. 5 (1961), 185–191.

*———, "Evaluation Research Programs in Public Health Practice," *Annals of the New York Academy of Sciences*, vol. 107, No. 2 (1963), 717–724.

*Fox, David J., "Issues in Evaluating Programs for Disadvantaged Children," *Urban Review*, vol. 2 (December 1967), 7, 9, 11.

*Freeman, Howard E. and Clarence C. Sherwood, "Research in Large-Scale Intervention Programs," *Journal of Social Issues*, vol. 21, No. 1 (1965), 11–28.

*Getting, Vlado A., "Part II—Evaluation," *American Journal of Public Health*, vol. 47, No. 4 (1957), 409–413.

*Glaser, Edward M. and Hubert S. Coffey, *Utilization of Applicable Research and Demonstration Results*. Los Angeles, California: Human Interaction Research Institute, n.d.

*Glennan, Thomas K., Jr., *Evaluating Federal Manpower Programs: Notes and Observations*. Santa Monica, California: The Rand Corporation, September 1969.

*Glock, Charles Y., et al., *Case Studies in Bringing Behavioral Science Into Use: Studies in the Utilization

of Behavioral Science*, vol. 1. Stanford: Institute for Communication Research, 1961.

Glock, Charles Y., (Ed.), *Survey Research in the Social Sciences*, New York: Russell-Sage Foundation, 1967.

*Gollin, Albert E., "The Evaluation of Overseas Programs: Applied Research and Its Organizational Context," in *Education and Training for International Living: Concepts*, edited by Robert Campbell, Bert King and John Nagay. Arlington, Virginia: Beatty Publishers, 1970.

*Gorham, William, "Notes of a Practitioner," *The Public Interest*, No. 8 (summer 1967), 4–8.

Greenberg, Bernard G. and Berwyn F. Mattison, "The Whys and Wherefores of Program Evaluation," *Canadian Journal of Public Health*, vol. 46, No. 7 (1955), 293–299.

*Griessman, B. Eugene, "An Approach to Evaluating Comprehensive Social Projects," *Educational Technology*, vol. 9, No. 2 (1969), 16–19.

Gruenberg, Ernest M. (ed.), "Evaluating the Effectiveness of Mental Health Services," *Milbank Memorial Fund Quarterly*, vol. 44, No. 1 (1966), part 2 (whole issue).

*Guba, Egon G., "Development, Diffusion and Evaluation," in *Knowledge Production and Utilization in Educational Administration*, edited by Terry L. Eidell and Joanne M. Kitchel, pp. 37–63. Eugene, Oregon: University Council for Educational Administration and Center for the Advanced Study of Educational Administration, University of Oregon, 1968.

*———, "The Failure of Educational Evaluation," *Educational Technology*, vol. 9, No. 5 (1969), 29–38.

*——— and John Horvat, "Evaluation During Development," *Bulletin of the School of Education, Indiana University*, vol. 46, No. 2 (1970), 21–45.

*Hagen, Elizabeth P. and Robert L. Thorndike, "Evaluation," in *Encyclopedia of Educational Research*, 3rd edition, pp. 482–486. New York: The Macmillan Co., 1960.

*Hall, Richard H., "The Applied Sociologist and Organizational Sociology," in *Sociology in Action*, edited by Arthur B. Shostak, pp. 33–38. Homewood, Illinois: Dorsey Press, 1966.

Halpert, Harold P., William J. Horvath, and John P. Young, *An administrators handbook on the application of operations research to the management of mental health systems*, Washington, D. C.: Public Health Service Publication No. 2110.

*Hardin, Einar and Michael E. Borus, "An Economic Evaluation of the Retraining Program in Michigan: Methodological Problems of Research," *Proceedings of the Social Statistics Section*, American Statistical Association (1966), 133–137.

*Hastings, J. Thomas, "Curriculum Evaluation: The Why of the Outcomes," *Journal of Educational Measurement*, vol. 3, No. 3 (1966), 27–32. Also reprinted in *Readings in Measurement and Evaluation*, edited by Norman Gronlund, pp. 53–60. New York: The Macmillan Co., 1968.

*Havelock, Ronald G., *Planning for Innovation Through

Dissemination and Utilization of Knowledge*. Ann Arbor: Institute for Social Research, University of Michigan, 1969.

Hayes, Samuel P., *Measuring the Results of Development Projects*. Paris: UNESCO, 1959.

*———, *Evaluating Development Projects: A Manual for the Use of Field Workers*. Paris: UNESCO, 1966.

*Hemphill, John K., "The Relationships Between Research and Evaluation Studies," in *Educational Evaluation: New Roles, New Means*, 68th Yearbook of the National Society for the Study of Education, edited by Ralph W. Tyler, pp. 189–220. Chicago: National Society for the Study of Education, 1969.

*Herman, Melvin, "Problems of Evaluation," *The American Child*, vol. 47, No. 2 (1965), 5–10.

*——— and Michael Munk, *Decision Making in Poverty Programs: Case Studies from Youth-Work Agencies*, pp. 139–181. New York: Columbia University Press, 1968.

*Herzog, Elizabeth, *Some Guide Lines for Evaluative Research*. Washington, D.C.: U.S. Department of Health, Education and Welfare, 1959.

*Hesseling, P., "Principles of Evaluation," *Social Compass*, vol. 11, No. 1 (1964), 5–22.

Hill, John G., "Cost Analysis of Social Work Service," In: Norman A. Polansky (Ed.), *Social Work Research*, Chicago: The University of Chicago Press, 1960.

*Hill, Marjorie J. and Howard T. Blane, "Evaluation of Psychotherapy with Alcoholics," *Quarterly Journal of Studies on Alcohol*, vol. 28, No. 1 (1967), 76–104.

*Holliday, L. P., *Appraising Selected Manpower Training Programs in the Los Angeles Area*. Santa Monica, Calif.: Rand Corp., May 1969.

*Hough, Robbin R., "Casualty Rates and the War on Poverty," *American Economic Review*, vol. 58, No. 2 (1968), 528–532.

*Hovland, Carl I., "Reconciling Conflicting Results Derived from Experimental and Survey Studies of Attitude Change," *American Psychologist*, vol. 14, No. 1 (1959), 8–17.

———, Arthur A. Lumsdaine and Fred D. Sheffield, *Experiments in Mass Communication*. Princeton, N.J.: Princeton University Press, 1949.

*Hutchison, George B., "Evaluation of Preventive Services," *Journal of Chronic Diseases*, vol. 11, No. 5 (1960), 497–508.

*Hyman, Herbert H. and Charles R. Wright, "Evaluating Social Action Programs," in *The Uses of Sociology*, edited by Paul F. Lazarsfeld, William H. Sewell, and Harold L. Wilensky, pp. 741–782. New York: Basic Books, Inc., 1967.

*James, George, "Planning and Evaluation of Health Programs," in *Administration of Community Health Services*, pp. 114–134. Chicago: International City Managers Association, 1961.

*———, "Research by Local Health Departments: Problems, Methods, Results," *American Journal of Public Health*, vol. 48, No. 3 (1958), 353–361.

Jenks, Charles L., "Evaluation for a Small District,"

Educational Product Report, vol. 2, No. 5 (1967), 8–17.

Jones, James A., "Research," in *Breakthrough for Disadvantaged Youth*, pp. 235–259. Washington, D.C.: U.S. Department of Labor, Manpower Administration, 1969.

*Justman, Joseph, "Problems of Researchers in Large School Systems," *Educational Forum*, Vol. 32, No. 4 (1968), 429–437.

*Kandel, Denise B. and Richard H. Williams, *Psychiatric Rehabilitation: Some Problems of Research*. New York: Atherton Press, 1964.

*Kelman, Howard R. and Jack Elinson, "Strategy and Tactics of Evaluating a Large Scale Medical Care Program," *Proceedings of the Social Statistics Section*, American Statistical Association (1968), 169–191.

Kiresuk, T. J., and R. E. Sherman, "Goal Attainment Scaling: A general method for evaluating comprehensive community mental health programs," *Community Mental Health Journal*, vol. 4, (1968), 443–453.

*Klineberg, Otto, "The Problem of Evaluation Research," *International Social Science Bulletin*, vol. 7, No. 3 (1955), 346–352.

*Kogan, Leonard S. and Ann W. Shyne, "Tender-Minded and Tough-Minded Approaches in Evaluative Research," *Welfare in Review*, vol. 4, No. 2 (1966), 12–17.

*Krause, Elliott A., "After the Rehabilitation Center," *Social Problems*, vol. 14, No. 2 (1966), 197–206.

*LaSorte, Michael A., "The Caseworker as Research Interviewer," *American Sociologist*, vol. 3, No. 3 (1968), 222–225.

*Lemkau, Paul V. and Benjamin Pasamanick, "Problems in Evaluation of Mental Health Programs," *American Journal of Orthopsychiatry*, vol. 27, No. 1 (1957), 55–58.

*Lempert, Richard, "Strategies of Research Design in the Legal Impact Study," *Law and Society Review*, vol. 1, No. 1 (1966), 111–132.

*Lerman, Paul, "Evaluative Studies of Institutions for Delinquents: Implications for Research and Social Policy," *Social Work*, vol. 12, No. 4 (1968), 55–64.

Levine, Abraham S., "Cost-benefit analysis in social welfare: An exploration of possible applications," *Welfare in review*, vol. 4, No. 2 (1966), 1–11.

*Levine, Abraham S., "Evaluating Program Effectiveness and Efficiency: Rationale and Description of Research in Progress," *Welfare in Review*, vol. 5, No. 2 (1967), 1–11.

*Levine, Robert A., "Evaluating the War on Poverty," in *On Fighting Poverty: Perspective from Experience*, edited by James L. Sundquist, pp. 188–216. New York: Basic Books, Inc., 1969.

*Levinson, Perry, "Evaluation of Social Welfare Programs: Two Research Models," *Welfare in Review*, vol. 4, No. 10 (1966), 5–12.

*Levitan, Sar, "Facts, Fancies, and Freeloaders in Evaluating Anti-Poverty Programs," *Poverty and Human Resources Abstracts*, vol. 4, No. 6 (1969), 13–16.

*Likert, Rensis, and Ronald Lippitt, "The Utilization of Social Science," in *Research Methods in the Behavioral Sciences*, edited by Leon Festinger and Daniel Katz, pp. 581–646. New York: Holt, Rinehart and Winston, 1953.

*Lippitt, Ronald, "The Use of Social Research to Improve Social Practice," *American Journal of Orthopsychiatry*, vol. 35, No. 3 (1965), 663–669.

*————, Jeanne Watson, and Bruce Westley, *The Dynamics of Planned Change*, pp. 263–272. New York: Harcourt Brace and Co., 1958.

*Longood, Robert and Arnold Simmel, "Organizational Resistance to Innovation Suggested by Research," in *Evaluating Action Programs*, edited by Carol H. Weiss. Boston: Allyn and Bacon, 1972.

*Luchterhand, Elmer, "Research and the Dilemmas in Developing Social Programs," in *The Uses of Sociology*, edited by Paul F. Lazarsfeld, William H. Sewell, and Harold L. Wilensky, pp. 513–517. New York: Basic Books, 1967.

*McDill, Edward L., Mary S. McDill, and J. Timothy Sprehe, *Strategies for Success in Compensatory Education: An Appraisal of Evaluation Research*. Baltimore, Md.: Johns Hopkins Press, 1969.

*McIntyre, Robert B. and Calvin C. Nelson, "Empirical Evaluation of Instructional Materials," *Educational Technology*, vol. 9, No. 2 (1969), 24–27.

*Mangum, Garth L., "Evaluating Manpower Programs," *Monthly Labor Review*, vol. 91, No. 2 (1968), 21–22.

*Mann, John, "The Outcome of Evaluative Research," in *Changing Human Behavior*, pp. 191–214. New York: Charles Scribner's Sons, 1965.

*Marris, Peter and Martin Rein, "Research," in *Dilemmas of Social Reform*, pp. 191–207. New York: Atherton Press, 1969.

*Mauldin, W. Parker and John A. Ross, "Family Planning Experiments: A Review of Design," *Proceedings of the Social Statistics Section*, American Statistical Association (1966), pp. 278–282.

*Merton, Robert K., "Role of the Intellectual in Public Bureaucracy," in *Social Theory and Social Structure* (1964 edition), pp. 207–234. Glencoe, Illinois: The Free Press, 1964.

*Meyer, Alan S. and Stanley K. Bigman, "Contextual Considerations in Evaluating Narcotic Addiction Control Programs," *Proceedings of the Social Statistics Section*, American Statistical Association (1968), pp. 175–180.

*Miller, S. M., "The Study of Man: Evaluating Action Programs," *Transaction*, vol. 2 (March/April, 1965), 38–39.

*Morehead, Mildred, "The Medical Audit as an Operational Tool," *American Journal of Public Health*, vol. 57, No. 9 (1967), 1643–1656.

*Moss, L., "The Evaluation of Fundamental Education," *International Social Science Bulletin*, vol. 7, No. 3 (1955), 398–417.

*Nagpaul, Hans, "The Development of Social Research in an Ad Hoc Community Welfare Organization," *Journal of Human Relations*, vol. 14, No. 4 (1966), 620–633.

*Ott, Jack M., "Classification System for Decision Situations: An Aid to Educational Planning and Evaluation," *Educational Technology*, vol. 9, No. 2 (1969), 20–23.

*Owens, Thomas R., "Suggested Tasks and Roles of Evaluation Specialists in Education," *Educational Technology*, vol. 8, No. 22 (1968), 4–10.

*Perry, S. E. and Lyman Wynne, "Role Conflict, Role Definition, and Social Change in a Clinical Research Organization," *Social Forces*, vol. 38, No. 1 (1959), 62–65.

*Provus, Malcolm, "Evaluation of Ongoing Programs in the Public School System," in *Educational Evaluation: New Roles, New Means*, 68th Yearbook of the National Society for the Study of Education, edited by Ralph W. Tyler, pp. 242–283. Chicago: National Society for the Study of Education, 1969.

Riecken, Henry W., "Memorandum on Program Evaluation," in *Evaluating Action Programs*, edited by Carol H. Weiss, Boston: Allyn and Bacon, 1972.

Riley, Matilda White, *Sociological research: A case approach*. New York: Harcourt, Brace & World, 1963.

*Rodman, Hyman and Ralph L. Kolodny, "Organizational Strains in the Researcher-Practitioner Relationship," in *Applied Sociology: Opportunities and Problems*, edited by Alvin Gouldner and S. M. Miller, pp. 93–113. New York: The Free Press, 1965.

*Rosenblatt, Aaron, "The Practitioner's Use and Evaluation of Research," *Social Work*, vol. 13, No. 1 (1968), 53–59.

*Rossi, Peter H., "Boobytraps and Pitfalls in the Evaluation of Social Action Programs," *Proceedings of the Social Statistics Section*, American Statistical Association (1966), pp. 127–132.

*————, "Evaluating Social Action Programs," *Transaction*, vol. 4, No. 7 (1967), 51–53.

*————, "Practice, Method, and Theory in Evaluating Social-Action Programs," *On Fighting Poverty: Perspectives from Experience*, edited by James L. Sundquist, pp. 217–234. New York: Basic Books, Inc., 1969.

*Sadofsky, Stanley, "Utilization of Evaluation Results: Feedback into the Action Program," in *Learning in Action*, edited by June L. Shmelzer, pp. 22–36. Washington, D.C.: Government Printing Office, 1966.

*Scanlon, R. G., "Innovation Dissemination," *Pennsylvania School Journal*, vol. 116 (March 1968), 375–376.

Schonfeld, H. F., I. S. Falk, P. H. Lavietes, I. Landwirth, and L. S. Krassnor, "The development of standards for the audit and planning of medical care," *American Journal of Public Health*, vol. 88, No. 11 (1968), 2097–2110.

*Schulberg, Herbert C. and Frank Baker, "Program Evaluation Models and the Implementation of Research Findings," *American Journal of Public Health*, vol. 58, No. 7 (1968), 1248–1255.

*————, Alan Sheldon and Frank Baker, *Program Evaluation in the Health Fields*. New York: Behavioral Publications, Inc., 1970.

Schwartz, Richard D., "Field Experimentation in Sociolegal Research," *Journal of Legal Education*, vol. 13, No. 3 (1961), 401–410.

*Scriven, Michael, "The Methodology of Evaluation," in *Perspectives of Curriculum Evaluation*, edited by Ralph W. Tyler, Robert M. Gagné and Michael Scriven (AERA monograph series on curriculum evaluation, No. 1), pp. 39–83. Chicago: Rand McNally, 1967.

*————, "An Introduction to Meta-Evaluation," *Educational Product Report*, vol. 2, No. 5 (1969), 36–38.

*Sheldon, Eleanor B. and Howard E. Freeman, "Notes on Social Indicators: Promises and Potential," *Policy Sciences*, vol. 1 (1970), 97–111.

*Sherwood, Clarence C., "Issues in Measuring Results of Action Programs," *Welfare in Review*, vol. 5, No. 7 (1967), 13–18.

*Slocum, W. L., "Sociological Research for Action Agencies: Some Guides and Hazards," *Rural Sociology*, vol. 21, No. 2 (1956), 196–199.

*Smith, Bruce L. R., *The Rand Corporation*, pp. 216–237. Cambridge, Masachusetts: Harvard University Press, 1966.

*Smith, Joel, Francis M. Sim, and Robert C. Bealer, "Client Structure and the Research Process," in *Human Organization Research*, edited by Richard N. Adams and Jack J. Preiss, chapter 4. Homewood, Illinois: Dorsey Press, 1960.

Somers, Gerald G., "Research Methodology in the Evaluation of Retraining Programmes," University of Wisconsin, Industrial Relations Research Institute, reprint series, No. 61. Reprinted from *Labour and Automation*, Bulletin No. 1, Geneva, 1965.

*Stake, Robert E., "Testing in the Evaluation of Curriculum Development," *Review of Educational Research*, vol. 38, No. 1 (1968), 77–84.

*————, "The Countenance of Educational Evaluation," *Teachers College Record*, vol. 68, no. 7 (1967), 523–540.

*————, "Generalizability of Program Evaluation: The Need for Limits," *Educational Product Report*, vol. 2, no. 5 (1969), 38–40.

Stein, Herman D., "The Study of Organizational Effectiveness," in *Research in Social Welfare Administration*, edited by David Fanshel, pp. 22–32. New York: National Association of Social Workers, 1962.

*————, George M. Hougham, Serapio R. Zalba, "Assessing Social Agency Effectiveness: A Goal Model," *Welfare in Review*, vol. 6, No. 2 (1968), 13–18.

Stember, Charles H., "Evaluating Effects of an Integrated Classroom," *The Urban Review*, vol. 2, No. 7 (1968), 3–4, 30, 31.

*Steward, M. A., "The Role and Function of Educational Research—I," *Educational Research*, vol. 9, No. 1 (1966), 3–6.

*Stouffer, Samuel A., "Some Observations on Study Design," *American Journal of Sociology*, vol. 55, No. 4 (1950), 355–361.

Stufflebeam, Daniel L., "The Use and Abuse of Evaluation in Title III," *Theory into Practice*, vol. 6, No. 3 (1967), 126–133.

*————, "Evaluation as Enlightenment for Decision-Making." An address delivered at the Working

Conference on Assessment Theory, sponsored by the Commission on Assessment of Education Outcomes, The Association for Supervision and Curriculum Development, Sarasota, Florida, Jan. 19, 1968. Columbus, Ohio: The Evaluation Center, College of Education, The Ohio State University.

———, "Toward a Science of Educational Evaluation," *Educational Technology*, vol. 8, No. 14 (1968), 5–12.

Sturz, Herbert, "Experiments in the Criminal Justice System," *Legal Aid Briefcase* (February, 1967), 1–5.

*Suchman, Edward A., "Action for What? A Critique of Evaluative Research," in *The Organization, Management, and Tactics of Social Research*, edited by Richard O'Toole. Cambridge, Massachusetts: Schenkman Publishing Company, 1970.

*———, *Evaluative Research: Principles and Practice in Public Service and Social Action Programs*. New York: Russell Sage Foundation, 1967.

*———, "A Model for Research and Evaluation on Rehabilitation," in *Sociology and Rehabilitation*, edited by Marvin B. Sussman, pp. 52–70. Washington, D.C.: American Sociological Association, 1966.

*Takishita, John Y., "Measuring the Effectiveness of a Family Planning Program: Taiwan's Experience," *Proceedings of the Social Statistics Section*, American Statistical Association (1966), pp. 268–271.

*Taylor, Philip H., "The Role and Function of Educational Research," *Educational Research*, vol. 9, No. 1 (1966), 11–15.

*Therkildsen, Paul and Philip Reno, "Cost-Benefit Evaluation of the Bernalillo County Work Experience Project," *Welfare in Review*, vol. 6, no. 2 (1968), 1–12.

Tripodi, Tony, Phillip Fellin, and Irwin Epstein, *Social Program Evaluation*, Itasca, Illinois: Peacock, 1971.

*Turvey, Ralph and A. R. Prest, "Cost-Benefit Analysis: A Survey," *Economic Journal*, vol. 75, no. 300 (1965), 683–735.

Tyler, Ralph W., "Assessing the Progress of Education," *Science Education*, vol. 50, No. 3 (1966), 239–242.

*———, Robert M. Gagné and Michael Scriven, *Perspectives of Curriculum Evaluation* (AERA monograph series on curriculum evaluation, No. 1). Chicago: Rand McNally, 1967.

*U. S. Congress, House Committee on Government Operations, Research and Technical Programs Subcommittee, *The Use of Social Research in Federal Domestic Programs*, vol. III. 90th Congress, 1st session. Washington, D. C.: Government Printing Office, 1967.

*U. S. Department of Health, Education, and Welfare, National Advisory Mental Health Council, *Evaluation in Mental Health*. Washington, D.C.: Public Health Service, Publication No. 413, 1955.

*U. S. Department of Health, Education and Welfare, Office of Education, *Preparing Evaluation Reports: A Guide for Authors*. Washington, D.C. Government Printing Office, 1970.

*Wall, W. D., "The Future of Educational Research,"

Educational Research, vol. 10, No. 3 (1968), 163–169.

Ward, David A. and Gene G. Kassebaum, "Evaluations of Correctional Treatment: Some Implications of Negative Findings," *Law Enforcement Science and Technology*, edited by S. A. Yefsky (proceedings of the First National Symposium on Law Enforcement Science and Technology, IIT Research Institute), pp. 201–209. Washington, D.C.: Thompson Book Co., 1967.

*Wardrop, James L., "Generalizability of Program Evaluation: The Danger of Limits," *Educational Product Report*, vol. 2, No. 5 (1969), 41–42.

*Weinberger, Martin, "Evaluating Educational Programs: Observations by a Market Researcher," Urban Review, vol. 3, No. 4 (1969), 23–26.

*Weiss, Carol H., *Evaluating Action Programs*, Boston: Allyn and Bacon. 1972.

*———, Evaluation of In-Service Training," in *Targets for In-Service Training*, pp. 47–54, Washington, D.C.: Joint Commission on Correctional Manpower and Training, October 1967.

*———, *Evaluation Research*, New York: Prentiss-Hall, 1972, In press.

*———, "Planning an Action Project Evaluation," in *Learning in Action*, edited by June L. Shmelzer, pp. 6–21. Washington, D.C.: Government Printing Office, 1966.

*———, "The Politicization of Evaluation Research," *Journal of Social Issues*, vol. 26, No. 4 (1970), 57–68.

———, "Prevention of Juvenile Delinquency: Research and Evaluation," in *Papers on Research in Crime and Delinquency*, pp. 1–21. Washington, D.C.: U.S. Department of Health, Education, and Welfare, Office of Juvenile Delinquency and Youth Development, December, 1966.

*———, "Utilization of Evaluation: Toward Comparative Study," in *The Use of Social Research in Federal Domestic Programs*, vol. III, pp. 426–432. U.S. Congress, House Committee on Government Operations, Research and Technical Programs Subcommittee, 90th Congress, 1st session, April 1967. Washington, D.C.: Government Printing Office, 1967.

*Weiss, Robert S. and Martin Rein, "The Evaluation of Broad-Aim Programs: A Cautionary Case and a Moral," *The Annals of the American Academy of Political and Social Science*, vol. 385 (September, 1969), 133–142. A revised version also appears in *Administrative Science Quarterly*, vol. 15, No. 1 (1970), 97–109.

Welty, Gordon A., "Experimental Designs and Applied Research," *California Journal of Educational Research*, vol. 20, No. 1 (1969), 40–44.

Western Interstate Commission for Higher Education, *Systems approach to program evaluation in mental health*, Boulder, Colorado: WICHE, 1970.

*Wholey, Joseph S., et al., *Federal Evaluation Policy*. Washington, D.C.: The Urban Institute, 1970.

*Wilder, David E., "Problems of Evaluation Research," in *An Overview of Adult Education Research*, edited by Edmund deS. Brunner, David E. Wilder, Corinne

Kirchner, and John S. Newberry, pp. 243–273. Chicago: Adult Education Association of the U.S.A., 1959.

*Wilkins, Leslie T., "Evaluation of Training Programs," in *Social Deviance*, pp. 288–293. London: Tavistock Publications, 1964.

*———, *Evaluation of Penal Measures*. New York: Random House, 1969.

Williams, Richard H., and Lucy D. Ozarin (Eds.) *Community Mental Health: An International Perspective*, San Francisco: Jossey-Bass, 1968.

*Williams, Walter, "Developing an Evaluation Strategy for a Social Action Agency," *Journal of Human Resources*, vol. 4, No. 4 (1969), 451–465.

*——— and John W. Evans, "The Politics of Evaluation: The Case of Head Start," *The Annals of the American Academy of Political and Social Science*, vol. 385 (September 1969), 118–132.

*Wittrock, M. C., "The Evaluation of Instruction," *Evaluation Comment* (Journal of the Center for the Study of Evaluation, UCLA), vol. 1, No. 4 (1969), 1–7. To be published in M. C. Wittrock and D. E. Wiley, *The Evaluation of Instruction: Issues and Problems* (in press).

*Worthen, Blaine R., "Toward a Taxonomy of Evaluation Designs," *Educational Technology*, vol. 8, No. 15 (1968), 3–9.

*Wright, Charles R. and Herbert H. Hyman, "The Evaluators," in *Sociologists at Work: Essays on the Craft of Social Research*, edited by Phillip E. Hammond, pp. 121–141. New York: Basic Books, 1964.

2. Illustrations of Evaluation Studies

*Aldrich, Nelson (ed.), "The Controversy Over the More Effective Schools: A Special Supplement," *Urban Review*, vol. 2, No. 6 (1968), 15–34.

*Benedict, Barbara A., Paula H. Calder, David M. Callahan, Harvey Hornstein and Matthew B. Miles, "The Clinical-Experimental Approach to Assessing Organizational Change Efforts," *Journal of Applied Behavioral Science*, vol. 3, No. 3 (1967), 347–380.

*Berleman, William C. and Thomas W. Steinburn, "The Execution and Evaluation of a Delinquency Prevention Program," *Social Problems*, vol. 14, No. 4 (1967), 413–423.

*Bogart, Leo, et al., *Social Research and the Desegregation of the U.S. Army*. Chicago: Markham Publishing Co., 1969.

*Boocock, Sarane S. and James S. Coleman, "Games with Simulated Environments in Learning," *Sociology of Education*, vol. 39, No. 3 (1966), 215–236.

*Cain, Glen and Gerald Somers, "Retraining the Disadvantaged Worker," in *Research in Vocational and Technical Education*, edited by Cathleen Quirk and Carol Sheehan, pp. 27–44. Madison: Center for Studies in Vocational and Technical Education, University of Wisconsin, 1967.

*——— and Ernst W. Stromsdorfer, "Retraining in West Virginia: An Economic Evaluation," in *Retraining the Unemployed*, edited by Gerald G. Somers, pp. 299–335. Madison: The University of Wisconsin Press, 1968.

*Campbell, Donald T. and H. Laurence Ross, "The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis," *Law and Society Review*, vol. 3, No. 1 (1968), 33–53.

*Caplan, Nathan, "Treatment Intervention and Reciprocal Interaction Effects," *Journal of Social Issues*, vol. 24, no. 1 (1968), 63–88.

*Clark, Burton R., *The Open Door College: A Case Study*. New York: McGraw Hill Book Company, Inc., 1960.

Chamberlin, C. D., Enid Chamberlin, N. E. Drought and W. E. Scott, *Adventure in American Education*, vol. IV, *Did They Succeed in College?* New York: Harper and Row, 1942.

*Coleman, James S., Ernest Q. Campbell, et al., *Equality of Educational Opportunity*. Washington, D.C.: Government Printing Office, 1966.

*Cumming, Elaine and John Cumming, *Closed Ranks: Study of Mental Health Education*. Cambridge, Mass.: Harvard University Press, 1957.

*Davis, James A., *Great Books and Small Groups*. Glencoe, Illinois: The Free Press, 1961.

Dentler, Robert, *The Young Volunteers: An Evaluation of Three Programs of the American Friends Service Committee*. Chicago: National Opinion Research Center, 1959.

*DiLorenzo, Louis and Ruth Salter, "An Evaluative Study of Prekindergarten Programs for Educationally Disadvantaged Children," *Exceptional Children*, vol. 35, No. 2 (1968), 111–120.

*Dressel, Paul L., *Evaluation in General Education*. Dubuque, Iowa: Wm. C. Brown Co., 1954.

Elkin, Robert, *Analyzing time, costs, and operations in a voluntary children's institution and agency*, Washington, D.C.: Department of Health, Education and Welfare, (September, 1965), 27–39.

*Fairweather, George W., *Social Psychology in Treating Mental Illness*. New York: John Wiley and Sons, 1964.

Gollin, Albert E., *Education for National Development: Effects of U.S. Technical Training Programs*. New York: Frederick A. Praeger, Inc., 1969.

Granger, R. L., et al., *The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development*, vol. I, Report to the Office of Economic Opportunity by Westinghouse Learning Corp. and Ohio University, June, 1969.

Greeley, Andrew M. and Peter H. Rossi, *The Education of Catholic Americans*. Chicago: Aldine Publishing Co., 1966.

*Hammond, K. R. and F. Kern, *Teaching Comprehensive Medical Care*. Cambridge, Massachusetts: Harvard University Press, 1959.

*Hyman, Herbert H., Charles R. Wright and Terence K. Hopkins, *Applications of Methods of Evaluation: Four Studies of the Encampment for Citizenship*. Los Angeles: University of California Press, 1962.

Jacob, Philip E., *Changing Values in College*. New York: Harper and Bros., 1957. (See Allen H. Barton, *Studying the Effects of a College Education*, listed in Section I.)

*Katz, Irwin, "Review of Evidence on Effects of De-segregation on the Intellectual Performance of Negroes," *American Psychologist*, vol. 19, No. 6 (1964), 381–399.

*Kellner, Robert, "The Evidence in Favour of Psychotherapy," *British Journal of Medical Psychology*, vol. 40, No. 4 (1967), 341–358.

*Kelman, H. R., "An Experiment in the Rehabilitation of Nursing Home Patients," *Public Health Reports*, No. 77 (April 1962), 356–366.

*Kendall, Patricia, "Evaluating an Experimental Program in Medical Education," in *Innovations in Education*, edited by Matthew B. Miles, pp. 343–360. New York: Teachers College, Bureau of Publications, 1964.

*Landers, Jacob, *Higher Horizons: Progress Report*. New York: Board of Education of the City of New York, January 1963.

*Lipton, Douglas, Robert Martinson, and Judith Wilks, *Treatment Evaluation Survey* (tentative title). State of New York (forthcoming).

*McCord, William and Joan McCord, *Origins of Crime: A New Evaluation of the Cambridge-Somerville Youth Study*. New York: Columbia University Press, 1959.

*McDill, Edward L., Mary S. McDill and J. Timothy Sprehe, *Strategies for Success in Compensatory Education: An Appraisal of Evaluation Research*. Baltimore: Johns Hopkins Press, 1969.

*Main, Earl D., "A Nationwide Evaluation of M.D.T.A. Institutional Job Training," *Journal of Human Resources*, vol. 3, No. 2 (1968), 159–170.

*Meyer, H. J. and E. F. Borgatta, *An Experiment in Mental Patient Rehabilitation*. New York: Russell Sage Foundation, 1959.

*———, E. F. Borgatta and W. C. Jones, *Girls at Vocational High*. New York: Russell Sage Foundation, 1965.

Miles, Matthew, "Changes During and Following Laboratory Training: A Clinical-Experimental Study," *Journal of Applied Behavioral Science*, vol. 1, No. 3 (1965), 215–242.

*Miller, Walter B., "The Impact of a 'Total-Community' Delinquency Control Project," *Social Problems*, vol. 10, No. 2 (1962), 168–191.

Newcomb, Theodore M., *Personality and Social Change*. New York: Dryden Press, 1957.

Office of Economic Opportunity, *Preliminary Results of the New Jersey Graduated Work Incentive Experiment*, Washington, D.C.: OEO, 1970.

———, *Further Preliminary Results of the New Jersey Work Incentive Experiment*, Washington, D.C.: OEO, 1971.

———, *Reports from the 100-City CAP Evaluation*, Washington, D.C.: OEO, 1970.

Pattison, E. Mansell, Ronald Coe, and Robert J. Rhodes, "Evaluation of Alcoholism Treatment: A Comparison of Three Facilities," *Archives of General Psychiatry*, vol. 20, No. 4 (1969), 478–488.

Powers, Edwin and Helen Witmer, *An Experiment in the Prevention of Juvenile Delinquency: The Cambridge-Somerville Youth Study*. New York: Columbia University Press, 1951.

Popham, W. J. and J. M. Sadnavitch, "Filmed Science Courses in the Public School: An Experimental Approach," *Science Education*, vol. 45, No. 4 (1961), 327–335.

Price, Bronson, *School Health Services: A Selective Review of Evaluative Studies*. Washington, D.C.: U.S. Department of Health, Education, and Welfare, Social Security Administration (Children's Bureau), 1957.

Riecken, Henry, *The Volunteer Work Camp: A Psychological Evaluation*. Cambridge, Massachusetts: Addison-Wesley, 1952.

Rosenthal, Robert and Lenore Jacobson, *Pygmalion in the Classroom*. New York: Holt, Rinehart and Winston, 1968.

Ross, H. Laurance, Donald T. Campbell and Gene V. Glass, "Determining the Social Effects of a Legal Reform: The British 'Breathalyser' Crackdown of 1967," *American Behavioral Scientist*, vol. 13, No. 4, Mar./Apr. (1970), 493–509.

*Sheldon, Alan, "An Evaluation of Psychiatric After-Care," *British Journal of Psychiatry*, vol. 110 (1964), 662–667.

*Silver, G., *Family Medical Care*. Cambridge: Harvard University Press, 1963.

Smith, E. R. and R. W. Tyler, *Appraising and Recording Student Progress*. New York: Harper and Row, 1942.

*Stromsdorfer, Ernst W., "Determinants of Economic Success in Retraining the Unemployed: The West Virginia Experience," *The Journal of Human Resources*, vol. 3, No. 2 (1968), 139–152.

Thomson, Captane P., and Norman W. Bell, "Evaluation of a Rural Community Mental Health Program," *Archives of General Psychiatry*, vol. 20, No. 4 (1969), 448–456.

*U. S. Department of Agriculture, Agricultural Marketing Service, Food Distribution Division, *The Food Stamp Program: An Initial Evaluation of the Pilot Project*. Washington, D.C.: April 1962.

U. S. Department of Health, Education, and Welfare, National Advisory Mental Health Council, *Evaluation in Mental Health*. Washington, D.C.: Public Health Service, publication No. 413, 1955.

*———, *A Bibliographic Index of Evaluation in Mental Health*, prepared by James K. Dent. Washington, D.C.: Public Health Service, Publication No. 1545, October 1966.

*Vanecko, James J., "Community Mobilization and Institutional Change," *Social Science Quarterly*, vol. 50, No. 3 (1969), 609–630.

*Wallace, D., *The Chemung County Research Demonstration with Dependent Multi-Problem Families*. New York: State Charities Aid Association, 1965.

Wallen, Norman E. and Robert M.W. Travers, "Analysis and Investigation of Teaching Methods," in *Handbook of Research on Teaching*, edited by N. L. Gage, chapter 10. Chicago: Rand McNally, 1963.

*Weeks, H. Ashley, *Youthful Offenders at Highfields*. Ann Arbor: University of Michigan Press, 1958.

Weisbrod, Burton A., "Preventing High School Dropouts," in *Measuring Benefits of Government Investments*, edited by Robert Dorfman, Washington, D.C.: Brookings Institution, 1965.

*Wilner, Daniel M., R. P. Walkley, T. C. Pinkerton, and M. Tayback, *The Housing Environment and Family Life*. Baltimore, Md.: Johns Hopkins Press, 1962.

*Wrightstone, J. Wayne, Samuel D. McClelland, Judith I. Krugman, Herbert Hoffman, Norman Tieman, and Linda Young, *Assessment of the Demonstration Guidance Project*. New York: Board of Education of the City of New York, Division of Research and Evaluation, n.d.

3. References on Design, Measurement, Sampling, and Analysis

Adams, Georgia S., *Measurement and Evaluation in Education, Psychology, and Guidance*. New York: Holt, Rinehart, and Winston, 1964.

*Barton, Allen H., *Organizational Measurement*. Princeton, N.J.: College Entrance Examination Board, 1961.

———, "Measuring the Values of Individuals," *Religious Education*, Supplement (July-August, 1962), pp. 62–97.

*Blalock, Hubert M., Jr., *Causal Inferences in Nonexperimental Research*, Chapel Hill: University of North Carolina Press, 1961.

*Bonjean, Charles M., Richard J. Hill, and S. Dale McLemore, *Sociological Measurement: An Inventory of Scales and Indices*. San Francisco: Chandler, 1967.

*Buros, Oscar (ed.), *Sixth Mental Measurements Yearbook*. Highland Park, New Jersey: Gryphon Press, 1965.

*Campbell, Donald T. and Julian Stanley, "Experimental and Quasi-Experimental Designs for Research on Teaching," in *Handbook of Research on Teaching*, edited by N. L. Gage, pp. 171–246. Chicago: Rand McNally, 1963. Also reprinted as *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally, 1969.

*——— and Donald W. Fiske, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin*, vol. 56, No. 2 (1959), 81–105.

*——— and Albert Erlebacher, "How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful," in *Compensatory Education: A National Debate*, edited by J. Hellmuth, vol. 3 of *The Disadvantaged Child*. New York: Brunner and Mazel, 1970.

*Cochran, William G., *Sampling Techniques*. New York: John Wiley & Sons, 1953.

Cox, David R., *Planning of Experiments*. New York: John Wiley & Sons, 1958.

Davis, James A., *Elementary Survey Analysis*. Englewood Cliffs: Prentice-Hall, 1971.

Duncan, Otis Dudley, "Path Analysis: Sociological Examples," *American Journal of Sociology*, vol. 72, No. 1 (1966) 1–16.

Gronlund, Norman (ed.), *Readings in Measurement and Evaluation*. New York: The Macmillan Co., 1968.

Group for the Advancement of Psychiatry, *Psychiatric Research and the Assessment of Change*, vol. 6, Report no. 63, New York, 1966.

*Hansen, Morris H., William N. Hurwitz and William G. Madow, *Sample Survey Methods and Theory*, vol. 1: *Method and Applications*. New York: John Wiley & Sons, 1953.

*Harris, Chester W. (ed.), *Problems in Measuring Change*. Madison: University of Wisconsin Press, 1963.

*Hochstim, Joseph R., "A Critical Comparison of Three Strategies of Collecting Data from Households," *Journal of the American Statistical Association*, vol. 62, No. 319 (1967), 976–989.

Hyman, Herbert, *Survey Design and Analysis: Principles, Cases, and Procedures*. Glencoe, Illinois: The Free Press, 1955.

*Kahn, Robert L. and Charles R. Cannell, *The Dynamics of Interviewing*. New York: John Wiley & Sons, 1957.

Kendall, Patricia L., "A Review of Indicators Used in 'The American Soldier,'" in *The Language of Social Research*, edited by Paul F. Lazarsfeld and Morris Rosenberg, pp. 37–39. Glencoe, Illinois: The Free Press, 1955.

Kish, Leslie, "Selection of the Sample," in *Research Methods in the Behavioral Sciences*, edited by Leon Festinger and Daniel Katz, pp. 175–239. New York: Dryden Press, 1953.

———, *Survey Sampling*. New York: John Wiley & Sons, 1965.

Michael, William G. and Newton S. Metfessel, "A Paradigm for Developing Valid Measurable Objectives in the Evaluation of Educational Programs in Colleges and Universities," *Educational and Psychological Measurement*, vol. 27, No. 2 (1967), 373–383.

*Miller, Delbert C., *Handbook of Research Design and Social Measurement*. New York: David McKay Co., 1964.

Monroe, J. and A. L. Finkner, *Handbook of Area Sampling*. New York: Chilton Company, Book Division, 1959.

Oppenheim, A. N., *Questionnaire Design and Attitude Measurement*. New York: Basic Books, 1966.

*Plutchik, R., S. R. Platman and R. R. Fieve, "Three Alternatives to the Double-Blind," *Archives of General Psychiatry*, vol. 20 (1969), 428–432.

Robinson, J. P., R. Athanasiou and Kendra B. Head, *Measures of Occupational Attitudes and Occupational Characteristics*. Ann Arbor: Survey Research Center, University of Michigan, 1967.

*———, Jerrold G. Rusk and Kendra B. Head, *Measures of Political Attitudes*. Ann Arbor: Survey Research Center, University of Michigan, 1968.

*——— and Phillip R. Shaver, *Measures of Social Psychological Attitudes*. Ann Arbor: Survey Research Center, University of Michigan, 1969.

Rosenberg, Morris, *The Logic of Survey Analysis*. New York: Basic Books, 1968.

Russell, B. M., et al., *World Handbook of Political and*

Social Indicators. New Haven and London: Yale University Press, 1964.

*Shaw, Marvin E. and Jack M. Wright, *Scales for the Measurement of Attitudes*. New York: McGraw-Hill, 1967.

*Sjoberg, Gideon and Roger Nett, *A Methodology for Social Research*. New York: Harper & Row, 1968.

Slonin, Morris James, *Sampling*. New York: Simon & Schuster, 1966.

*Stephan, Frederick F. and Phillip J. McCarthy, *Sampling Opinions; An Analysis of Survey Procedure*. New York: John Wiley & Sons, 1958.

*Sudman, Seymour, *Reducing the Cost of Surveys*. Chicago: Aldine Publishing Co., 1967.

*Thorndike, Robert L. and Elizabeth Hagen, *Measurement and Evaluation in Psychology and Education*, 2d edition. New York: John Wiley & Sons, 1961.

*Underhill, Ralph, *Methods in the Evaluation of Pro-grams for Poor Youth*. Chicago: National Opinion Research Center, June 1968.

*U. S. Bureau of the Budget, Executive Office of the President, *Houshold Survey Manual*, 1969. (Available from National Technical Information Service, U.S. Dept. of Commerce, Springfield, Va. 22151, document #PB 18 7444.)

*Webb, Eugene J., Donald T. Campbell, R. D. Schwartz and L. B. Sechrest, *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally, 1966.

Weiss, Carol H., "Interviewing Low-Income Respondents," *Welfare in Review*, vol. 4, No. 8 (1966), 1–9.

Winer, B. J., *Statistical Principles in Experimental Design*. New York: McGraw-Hill Book Co., Inc., 1962.

Zeisel, Hans, "Reducing the Hazards of Human Experiments through Modification in Research Design," *Annals of the New York Academy of Sciences*, vol. 169 (1970), 475–486.

PART III

ABSTRACTS OF WORKS ON EVALUATION RESEARCH

Aldrich, Nelson (Ed.), "The Controversy Over the More Effective Schools: A Special Supplement," *Urban Review*, vol. 2, No. 6, (1968), 15–34.

This Special Supplement presents the conclusions of David Fox's 1966–67 evaluation of the More Effective Schools program in New York City, two critiques of Fox's work, and Fox's answer to the critics. The conclusions which Fox drew from the study of the MES program after 3 years of operation were: (1) there was great variation from one ME school to another, (2) the climate and attitude of the school, staff, and parents were characterized by hope and enthusiasm, no mean accomplishment in school-community relations, (3) the MES made no significant difference in the achievement of the children, and (4) there was no innovation in the basic teaching process.

David Schwager, Vice President of the United Federation of Teachers, which strongly supported the program, and Harry Gottesfeld, a psychology professor, criticize the evaluation on technical grounds: the failure to use urban norms for assessing the performance of urban children, poor use of control groups, failure to carry out longitudinal study, misinterpretation of data based on preconceived biases, lack of standardization in observers' reports, unreliability of measuring devices, misadministration of tests, biased teacher sampling, and insufficient statistical analysis. Both critics indicate their concern that the Board of Education was using the evaluation to phase out the MES program.

Fox replied that his evaluation was in fact short-term and thus merely suggestive. The program was too young to realize its potential and the Board would be misusing the evaluation if it chose to find only negative things about MES in the report. He defended his research techniques, granting that the data were open to differing interpretation, and reaffirmed his conclusions that there were both great strengths and weaknesses in the MES and that further study was needed before any definitive conclusions should be drawn or decisive action implemented.

Alkin, Marvin C. "Evaluation Theory Development," *Evaluation Comment*, vol. 2, No. 1 (1969), 2–7.

A model of evaluation research, based on the concept of evaluation areas, is defined and developed. This model is geared toward handling analysis both of total systems and of specific programs.

Previous definitions of evaluation have proved inadequate in describing the broad, complex activities required of the evaluator. The author's definition is based on the view that evaluation situations are often unique, and that the purpose of evaluation is to provide useful information to the decision-maker.

Five areas of evaluation are identified: (1) evaluations which provide information about the state of the system (systems assessment), (2) evaluations which provide information relevant to the selection of programs to serve specific educational needs (program planning), (3) evaluations which provide information about the extent to which a program has been introduced in the intended manner (program implementation), (4) evaluations which help in improving the program (program modification), and (5) evaluations which help the decision-maker to judge the overall worth of a program (program certification). Each of these areas requires different approaches and methodologies.

At the micro-level, the most important task is specification of objectives, while at the macro-level the social and organizational context of the system is the most crucial factor to be examined.

American Institutes for Research, *Evaluative Research: Strategies and Methods*. Pittsburgh: American Institutes for Research, 1970.
Hawkridge, David G. "Designs for Evaluative Studies," pp. 24–47.

The paper discusses the differences between the scientific and intuitive approaches to evaluation. The type of evaluation considered is summative evaluation, which is designed to help

decision-makers choose among alternative programs in the field of education.

In each of seven phases of evaluation, the analytic and intuitive schools take different approaches: (1) setting objectives for evaluation, (2) selecting objectives to be measured, (3) choosing instruments and procedures, (4) selecting samples and control groups, (5) establishing schedules for evaluation, (6) choosing analysis techniques (where the battle between the two approaches is fiercest), (7) drawing conclusions and making recommendations.

Unsatisfactory designs are formulated by both groups: the scientists may discount the realities of politics, human opinions, and circumstances, and the intuitive approach may neglect possible biases, use inaccurate instruments, or interpret on the basis of insufficient data. The author concludes that the analytical approach is more effective and alone can advance development in the field. However, the scientific group will have to retain a sense of realism and provide important, practical data.

Bend, Emil, "The Impact of the Social Setting Upon Evaluative Research," pp. 109–129.

This paper discusses recurring administrative problems in planning and conducting evaluations of social action programs. The evaluative research process is divided into three phases, and problems are discussed in each phase.

(1) The Planning and Preparation Phase. Problem: Inadequate information and differential expectations of sponsors and subjects can cause a severe gap between the researcher and divergent interests within the organization, causing a lack of understanding and agreement about evaluation objectives and activities. Problem: A lack of coordination within and between the organization and the research team can have unfortunate scheduling and attitudinal consequences.

(2) The "On-Site" Phase. Problem: Lack of acceptance of the research team by the subjects can result in incomplete, incorrect, and biased data. Problem: An evaluation project can be jeopardized by the failure of program staff to meet conditions imposed by the research design.

(3) The Analysis and Reporting Phase. Problem: Evaluative research findings and recommendations are often presented in a form

that makes them difficult for sponsors to interpret and apply.

Some suggestions are offered on how to minimize the unwanted effects of the environment on evaluation research. However, many reactions from the environment are justified and serve to instruct evaluators about inadequacies of applied social research.

Andrew, Gwen, "Some Observations on Management Problems in Applied Social Research," The American Sociologist, vol. 2, No. 2 (1967), 84–89, 92.

This paper discusses problems that arise in applied research, particularly when a separate research-demonstration is installed within an ongoing agency. Two primary sources of stress are: (1) the managerial arrangements of the project and (2) the professional strain between program and research demands. Drawing on the experience of four applied research programs, Andrew identifies such managerial problems as location of budgetary control, earmarked funds, identification of the project as an independent unit, recruitment of specialized personnel, tendency to use "expert" consultants, time lag in implementing the action program, conflict between levels of the organizational hierarchy, hostile interdisciplinary relationships, and isolated program decisions. Professional problems include limited use of available knowledge, limited definition of the program in professional terms, failure to specify conditions under which the program is expected to operate, limited specification of procedural rules, and the threat of failure. The author makes suggestions for resolution of the conflicts, including extended formal contact between researchers and the personnel department to arrange recruiting procedures, decisions on staff requirements well in advance of hiring, selection of local agency staff as part of the project staff and half-time appointments in the agency and on the project, program goals determined by program staff, incorporation of agency people in the research unit, staff knowledge of the literature, planning of intervention in terms of multiple decision processes in adaptive systems, concentrating on the processes by which goals are to be reached. Such procedures can decrease distrust on the part of practitioners and aid in creating well-designed evaluation studies.

Aronson, Sidney, and Clarence C. Sherwood, "Researcher versus Practitioner: Problems in Social Action Research," Social Work, vol. 12, No. 4 (1967), 89–96.

This article describes difficulties of evaluating federal programs directed toward delinquency control and the establishment of multi-service community centers. Problems occurred on all levels of contact between researcher and practitioner.

On the highest level, research directors found it difficult to get program designers to think in terms of goals and to conceptualize the intended steps from input to outcome. Constant probing was necessary to get away from processes to principles. Lower level practitioners, such as social workers, teachers, camp counselors, etc., did not appreciate the necessity for control group designs. Thus, even when control groups were set up by the program designer, the practitioners often "sabotaged" the requirements. There was also difficulty at the lower levels in obtaining the kinds of client-record data which were needed.

The researchers found some of the project directors hostile to research, since they had not been included in its design. Furthermore, they tended to change the programs constantly since they had little stake in testing the specified intervention mechanisms. In one case, one of the major innovations—the "anchor man" who would be the permanent liaison between the client and the multi-service center—was abandoned by several centers. Thus, the evaluation for these projects became irrelevant, often without any forewarning to the researcher.

The authors end on a hopeful note, indicating that these experiences taught the researchers, many of whom were from academic backgrounds, a great deal that could be applied in future evaluation studies.

Barton, Allen H. Organizational Measurement. Princeton, New Jersey: College Entrance Examination Board, 1961.

This book presents a wide variety of measures and classifications of organizations developed by researchers in many different fields. Types of variables and measures covered include: measures of input, measures of output, environmental variables, social structural variables, attitudes, and activities. Discussion

focuses on the logical nature of the measurement operation. An appendix gives a summary of empirical measures of college characteristics.

Barton, Allen H. Studying the Effects of College Education. New Haven: The Edward W. Hazen Foundation, 1959.

This monograph is a critique of Phillip E. Jacob's book, Changing Values in College. Jacob attempted to examine the results of a large number of studies on changing values and draw general conclusions from them. Barton returns to the original studies, as well as to Jacob, and on methodological grounds, he questions their comparability and their relevance to the problem of values.

He begins with a review of what different researchers mean by values and makes explicit the underlying model of factors which influence behavior. Often no distinctions are made between values and their major determinants: the capacity for critical thought; the knowledge and beliefs on which this is based; and the degree of emotional sensitivity which the individual has acquired. The untested model, and the variability of definitions of values between researchers, make it difficult to summarize conclusively the results of past research.

Some of the major issues surrounding the problem of measuring values and beliefs are discussed, such as what is to measured, how it may be validly measured, etc. It is concluded that good specific measuring devices have yet to be developed. The usefulness of general types of measures is assessed.

Different types of design used in studies are compared. Because the college population is self-selected, it is impossible to follow the experimental model. However, some controls may be maintained by using before and after measures, comparisons of groups at different levels of exposure, measuring pre-existing differences between groups exposing themselves to different stimuli, and making a gross check on the alternative hypothesis of maturation or historical effects by comparisons with a "control group." Such controls are necessary if it is to be proved that college, rather than other factors, produced changes in values.

The last chapter discusses problems of specification and generalization. Colleges are complex social systems, and proving that they affect

values gives little information about what parts of the college experience may be relevant. Relations between subsystems must be analyzed if sources of influence or barriers to influence are to be determined. Furthermore, different colleges may exhibit different patterns of subsystem influence, and different types of students may react differently to the same stimuli. Because the efforts to measure value change in college have been scattered and short-range, there is some question about whether any broad generalizations may be drawn from them. Jacob's overall conclusions from the data—that college has little effect on values—may therefore be better seen as a hypothesis deserving of more systematic research.

Bateman, Worth, "Assessing Program Effectiveness: A Rating System for Identifying Relative Project Success," *Welfare in Review*, vol. 6, No. 1 (1968), 1–10.

This paper recognizes that the over-all evaluation of a federal program can mask significant differences in effectiveness among individual projects within the same program. If federal funds are to be allocated wisely within a program, it is necessary to identify those local projects which are poorly managed and ineffective. However, it is not only poor management that can lead to disappointing outcome results; outside conditions (unemployment rates, participant characteristics, political atmosphere, etc.) can affect outcomes.

Bateman uses the Work Experience and Training Program (funded under Title V of the Equal Opportunity Act) as an example. The goal of the program was to increase the earning power of the unemployed poor through basic education, training and services. Four effectiveness measures were chosen: employment rate of project participants, their occupational distribution, average wage, and proportion who went on to further training. The author presents a procedure for controlling for the effect of outside conditions to determine the relative success of local projects. Projects were stratified into 48 categories on the basis of factors that regression analysis identified as important—local unemployment rate, percentage of trainees who were male, and average age of trainees. Each project was compared with others in the same category on each effectiveness measure and an

aggregate score was compiled. The scores were used in making decisions on which local projects should be renewed, phased out, or modified.

Further research should be done on the causes of variation within projects. [Ed. note: It might be rewarding to identify characteristics of projects across all categories that ranked high (or low) on effectiveness.]

Belshaw, Cyril, "Evaluation of Technical Assistance as a Contribution to Development," *International Development Review*, vol. 8 (June 1966), 2–23.

Evaluations of the effectiveness of technical aid to under-developed countries must be concerned not only with specific goal-achievement, but even more importantly with the development of skills, side effects, and other intangibles, and the assessment of the appropriateness of the goal itself. The important question which must be answered is whether the programs have assisted in the overall socio-economic development of a country. This question involves placing the program in the organic framework of the economic and social environment of the country. Single-criterion evaluations are unlikely to provide all the needed information to assess the impact of a program, since intervening variables and hidden factors, such as under-utilization of skills, will be ignored.

Belshaw offers a list of potentially useful criterion measures for evaluating technical assistance programs, such as alterations of demand and consumption patterns, emergence of new demands, changes in the division of labor, and creation of indigenous institutions and organizations to train people to produce and administer innovations. He calls for the development of better theory on the process of development, which can then be used in evaluation studies. Any assessment of the contribution of technical assistance will involve the specification of assumptions about the strategy of development in the particular circumstances of the country.

Benedict, Barbara A., Paula H. Calder, Daniel M. Callahan, Harvey A. Hornstein, and Matthew B. Miles, "The Clinical-Experimental Approach to Assessing Organizational Change Efforts," *Journal of Applied Be-*

havioral Science, vol. 3, No. 3 (1967), 347–379.

This article reports the attempt to institute and evaluate a T-group type of program designed to improve relations among principals and the superintendent of a school system. The innovative methodological approach was to combine the rigor of experimental design and measurement with the richness of the clinical approach to organizational intervention. The diagnostic/training staff, which planned and executed the intervention, was kept separate from the research staff, which developed and analyzed the instruments to measure change. This procedure helped to control for biases resulting from personal involvement on the part of the change staff and guarded against the feedback of information about results of measurement which might have contaminated the intervention procedures. A multiple-time series design was used which involved taking two measurements before and after the intervention on the experimental group and a control school staff. Specific hypotheses about the particular school system and general hypotheses about organizational reactions to clinical interventions were tested. Analysis showed that there were no significant changes as a result of the program, and several situation-specific hypotheses to account for this result are proposed.

Bennis, Warren G. "Theory and Method in Applying Behavioral Science to Planned Organizational Change," *Journal of Applied Behavioral Science*, vol. 1, No. 4 (1965), 337–360.

The article is of interest to the evaluator because it suggests the variety of approaches which the social scientist may take to planned change, and the number of possible variations in the degree of his involvement with the program under scrutiny. Eight different methods for bringing about change are identified: (1) exposition and propagation of the "truth," under the assumption that once the men in power have the "truth" they will follow it; (2) elite corps programs, which focus on attempts to get academics, scientists, social workers, etc., into the ruling elite; (3) human relations training programs; (4) staff programs, where the social scientist observes, analyzes, and helps to plan rationally; (5) scholarly consultation—ex-

ploratory inquiry, scholarly understanding, confrontation, discovery of solutions, and scientific advice; (6) circulation of ideas to the elite—getting direct access to the ear of the powerful; (7) developmental research—seeing whether an idea can be brought to an engineering stage [Ed. note: formative evaluations would fall into this category.]; (8) action research, which undertakes to solve a problem from planning stages through implementation.

In contrast with most discussions of the evaluator's role, which recommend that he limit himself to change programs of the types 5, 6, and 7, Bennis suggests that planned change programs will be most effective when the social scientist also takes on the role of change agent, becoming involved, at least to some degree, in the planning and implementation stages.

Berleman, William C. and Thomas W. Steinburn, "The Execution and Evaluation of a Delinquency Prevention Program," *Social Problems*, vol. 14, No. 4 (1967), 413–423.

This article reports the design and results of an experimental pre-test of a delinquency prevention program in the community. Because the pre-test study was severely circumscribed in time and population serviced, it did not represent a test of service. Rather it offered the opportunity for implementation of rigorous procedural and evaluative techniques. The study is notable for its attempts to record precisely both the selection procedures used for experimental and control groups, and the exact amount of service received during the pre-test period.

High Risk boys were chosen from a base population of 167 Negro boys in Seattle. Four types of High Risks were selected on the basis of previous behavior in school and community and predicted behavior on factors associated with acting out in previous age cohorts. Within each High Risk category boys were randomly assigned to control or experimental groups. Among the Highest Risks, a larger proportion were assigned to the experimental section, so that their expected higher dropout rate would not leave the experimental group with boys who were not as "bad" as the controls.

Four male social workers gave intensive service to the experimental group. After the initial dropouts, each had a group of seven to

work with. They met weekly at a community center, and the social worker also made contacts with them at school and at home. Extensive recordings were kept of every contact. The median amount of service time a boy and his significant others received during the pre-test was slightly in excess of 75 hours.

Data from school and court records were analyzed and weighted for severity of infractions. Data for each group—experimental, control, attrition, and Low Risk—were analyzed for four time periods before and after the service period. The data indicate that acting out was reduced for the experimental group during the period of service, but regressed to previous levels after the termination of the demonstration period.

Blalock, Hubert M., Jr. *Causal Inferences in Nonexperimental Research*. Chapel Hill: University of North Carolina Press, 1961.

This book suggests an approach to causal inference in nonexperimental research. Rather than focusing on the design of experimental controls in the research, the author presents methods of data analysis (suitable for any type of empirical data, including survey) which test alternative causal models, using the statistical concept of degrees of fit. For example, with three variables, there are several possible models of causal relationships—eg. A→ B→ C, A→ C→ B, etc. The alternatives are tested, and those which are less good in their fit with the data are discarded. More elaborate models can be cumulatively developed by adding variables. A theoretical presentation of the procedure is made, and two applications are given using actual data. The logic of the approach is developed in non-technical language which can be understood by a reader with a minimal background in mathematics; some statistical competence is necessary to apply the procedure.

Blenkner, Margaret, "Obstacles to Evaluative Research in Casework: Part I and Part II," *Social Casework*, vol. 31, Nos. 2 and 3 (1950), 54–60, 97–105.

This article discusses the pressing need for scientific evaluative research in social casework and the psychological, social, economic, and methodological obstacles to such research.

The mentality of a caseworker, warm and empathetic, is often very different from the conceptual analytic mentality of the scientific researcher. This often causes hostility between researchers and caseworkers, which prevents research from being done. Caseworkers claim that outsiders cannot understand the profession well enough to do valid research, yet they themselves seldom have the proper training. In addition, they are often afraid that funds will be cut if their results are negative. Good research is very expensive in time and money, and caseworkers are often overburdened with work, while their agency budgets can seldom support a research program. The research that has been done is often superficial.

Training in research methods is critical to caseworkers in order to learn reliable, valid measuring techniques in determining both final case results and the nature of the worker-client relationship. Research teams of both caseworkers and research technicians are needed to do the urgently needed research in the casework field.

Bogart, Leo, et al. *Social Research and the Desegregation of the U. S. Army*. Chicago: Markham Publishing Company, 1969.

This volume presents the results of the two major troop opinion surveys on desegregation conducted in 1951 for the U. S. Army, one in Korea and the second in the continental U. S. The studies were long classified by the Army as "Secret," and publication became possible only recently.

The large-scale questionnaire and interview surveys reported here were the main part of Project Clear, undertaken to study the possible effects of desegregation of the Army. Because 1950 military regulations established a policy of "equality of opportunity," which permitted but did not require integration, the researchers were able to compare attitudes of Negroes and Whites in both segregated and integrated outfits. They focused particularly on the analysis of opinions, morale, and race relations under differing conditions. Their results indicated that desegregation "works" and that the more experience men had in desegregated situations, the more favorable they were to desegregation.

Bogart's introduction to the volume describes the history of the surveys. While desegregation would have eventually taken place anyway because of ideological, political, and manpower pressures, the studies supplied clear support for a policy of total desegregation. Contemporary applications of Project Clear results are also suggested.

Bonjean, Charles M., Richard J. Hill and S. Dale McLemore, *Sociological Measurement: An Inventory of Scales and Indices*. San Francisco: Chandler Publishing Company, 1967.

The evaluator who is interested in locating useful scales and indices can turn to this book for a classification of available sociological measures. The authors present 78 major conceptual categories and list significant variables within each category. For example, one conceptual class is "Characteristics of Complex Organizations," and variables under this heading include administrative rationality, complexity, division of labor, and so on. For each variable listed, there is a bibliography of articles which have used relevant scales and indices; when the measure is one which has been used frequently, there is a description of it as well. (Bibliographies were obtained from a content analysis of four major sociological journals from 1954 to 1965.) This book thus helps the researcher not only to find existing measures but also guides him to previous studies relevant to the conceptual issues which he is facing.

Boocock, Sarane S. and James S. Coleman, "Games with Simulated Environments in Learning," *Sociology of Education*, vol. 39, No. 3 (1966), 215–236.

The use of games with simulated environments as a means of teaching teen-aged youth was evaluated in terms of enjoyment and specific learnings. Three games (career, legislative, and community disaster) were tested on 4-H Club members. Youth were randomly assigned to the career and legislative games, and each group became the control for the other. There was no control group for the community disaster game. Participants filled out questionnaires before and after the game session.

Enthusiasm for the games was high. Each game communicated specific information and concepts (e.g. in the legislative game, the need to build up exchange relationships in order to reach collective decisions). Three kinds of general learnings seem to have occurred in all games: an appreciation of the complexity of real-life situations, a feeling of greater control over the environment, and a sense of the interdependence in the environment.

Borgatta, Edgar F. "Research Problems in Evaluation of Health Service Demonstrations," *Millbank Memorial Fund Quarterly*, vol. 44, No. 4 (1966), part 2, 182–199.

Evaluation research requires a clear statement of what is to be evaluated. Although medical programs appear quite clear in intent, the addition of goals other than improved medical health complicates the situation. Priorities must be determined, decisions made on the primacy of short- or long-range effects, and criteria of success must be developed. In almost all cases, some program goals will be in conflict. In public health programs, one treatment may not be most effective for all subgroups. There may also be conflicts between goals and values, e.g., use of abortion in family planning.

In general, experimental design is the most rigorous and scientific method of evaluation, but it will not eliminate all problems of validity. Improper sampling, regression, pre-test interference with the treatment and many other factors may contaminate the situation. Looser designs are sometimes appropriate. For example, when a new drug is introduced there is often no need for a control group, but care must be taken in the analysis to examine alternative hypotheses to explain the change before conclusions are drawn. In demonstration projects in particular, staff enthusiasm about a new treatment may affect the outcome, and results may not persist over time or transfer to another population.

The evaluator must also be aware of outside factors, such as the feasibility of his analysis and recommendations in the light of public sentiments. He must evaluate in terms of culture and values as well as efficiency.

Borgatta, Edgar. "Research: Pure and Applied," *Group Psychotherapy*, vol. 8, No. 3 (1955), 263–277.

The author advocates the evaluation of the effectiveness of psychotherapy, but notes that

there is often resistance to such research among clinicians. Arguments frequently used to deny the need for such research are listed and rebutted. 1) If control groups are used, some of the persons who could profit most from the therapy may be in the control group. But this should be the case, since it is necessary for the purposes of experimentation that individuals of all types be in both the treated and untreated groups. 2) It is not fair to withhold treatment from some individuals by placing them in a control group. But if the therapy is not of value, it will be wasteful to treat them. 3) Experiments cannot answer the problems of the therapist. But even if no one experiment can do this, eventually the growing body of knowledge will become relevant to clinicians' day-to-day problems. 4) Experiments are not necessary, because the clinician knows what he is doing and is doing his best. This is not necessarily true. The clinician may have false impressions about what he is doing. 5) If therapies were not good, they would not be used. But popularity does not equal goodness, or therapists would also have to admit that witch doctors and tea-leaf readers were "good."

Brooks, Michael P. "The Community Action Program as a Setting for Applied Research," *Journal of Social Issues*, Vol. 21, No. 1 (1965), pp. 29–40.

This article discusses the possible uses of evaluation research in community action programs for the advantage of both the programs and the researcher. The researcher can provide ideas for experimentation in action programs, collect and analyze data necessary for program planning, and assist in the planning process by encouraging rationality. He can also design and implement the necessary and complex evaluation studies. Evaluation serves several purposes for community programs: (1) it informs the funding agent of the value received for dollars spent, (2) refines and improves the program through continuous feedback, (3) makes results of the program available to other interested communities, and (4) examines the underlying hypotheses of the program. The evaluation may be done on the level of individual projects such as pre-school centers, of the community program as a whole and its impact, or of the aggregate impact of a number of community

action programs. Two possible foci of the evaluation are the program product and the program process, although the two are closely interrelated.

Several constraints exist which may hinder effective evaluation. First is the long-standing tension between those involved in action and those who evaluate. Second is the constraint imposed by the disciplinary boundaries which separate the social sciences. Communities are obviously interdisciplinary. Third is the ethical necessity for continuous feedback of research results into the programs. This produces a changing program needing a changing research design. Fourth there is always pressure for immediate results. Fifth is the diversity and unpredictability of people involved in the community program.

Brunner, Edmund deS. "Evaluation Research in Adult Education," *International Review of Community Development*, No. 17–18 (1967), 97–102.

This article stresses the need for evaluation of adult education programs in terms of the participants as well as the administrators. The assumption is usually made that the objectives of the participants are the same as those of the educators, but many diverse needs and interests have been found to motivate adults to participate in these programs. The evaluation design might include a scale to measure the satisfaction of participants (controlling for sex, educational, socioeconomic and ethnic status, since all these factors have been shown to affect adult education programs). Some of the best research has been done by the Extension Service of the U. S. Department of Agriculture. The Service has, for example, studied the diffusion and adoption of new ideas in agriculture and home economics to reveal the type of persons most likely to innovate.

Buros, Oscar K. (Ed.), *The Sixth Mental Measurements Yearbook*. Highland Park, New Jersey: The Gryphon Press, 1965.

The first section of this volume provides reviews of achievement and aptitude tests by experts in psychometrics, along with information on the subject matter, purposes, administration and scoring procedures, references to research in which the test has been used, and

reliability and validity information. The tests cover such areas as academic achievement (history, mathematics, etc.), general intelligence and aptitude, character and personality measures, business and clerical skills, dexterity and coordination, etc. The second section is a classified index of books and book reviews on testing.

Bynder, Herbert, "Sociology in a Hospital: A Case Study in Frustration," in *Sociology in Action*, edited by Arthur B. Shostak, pp. 61–70. Homewood, Illinois: Dorsey Press, 1966.

The author, who was hired by the social service agency of a hospital to conduct both evaluative and pure research projects, documents the difficulties of working within a bureaucratic organization. From the beginning he encountered hostility from the social-work oriented staff, who believed that he should not be trusted to do research vital to the department's concerns until he had shown his commitment to their values. The expectation was that he would fill the role of an agency social worker. Attempts were made to gain administrative control over the research process, to determine not only the general areas of inquiry, but also the specific topics, methodologies, and framework of the findings. The friction stemmed in part from the marginal position of the social agency in the hospital, and their felt need to justify their usefulness. Outside of the social service agency, clinical staff frustrated attempts to conduct meaningful research. Their treatment perspective and their greater importance in the hospital system allowed them to reject important proposals. Finally, the author found himself becoming overly committed to social action as compared to social scientific goals.

Cain, Glen G. and Robinson G. Hollister, *The Methodology of Evaluating Social Action Programs*. Madison: The Institute for Research on Poverty, University of Wisconsin, n.d. (*Unpublished*)

This paper examines evaluation methodology in the context of social action programs, particularly anti-poverty programs. Evaluations should provide a model suitable for statistical testing, a wide range in the values of variables representing program inputs, and judicious use of control groups. Two types of evaluation are identified: process and outcome. Process evalua-

tion maintains a check of the honesty of financial transactions as well as other managerial functions. Outcome evaluation is cost-benefit analysis involving an evaluation of the entire program concept.

Requirements of evaluation design include specification of program objectives, use of control groups, agreement on the level of decision-making at which the results are to be used, providing a statistical model for testing, applying economic theories, and timing the evaluation so that enough information is available. Because of the present state of social behavior theories, a single pilot project does not provide much information about alternative courses of action. A combination of loose administration, rapid operational changes in individual projects, and large-scale programs with heterogeneous projects creates natural experiments for evaluation design. However, intentional experiments, where different models are used in different projects would be much more satisfactory. The investigator must be committed to holding to a design for a long enough period of time to learn something useful. Because social action programs are so complex, a judgment of success or failure is too simple, and evaluation evidence should be used to suggest modifications.

Cain, Glen G. and Ernst W. Stromsdorfer, "Retraining in West Virginia: An Economic Evaluation" in *Retraining the Unemployed*, edited by Gerald G. Somers, pp. 299–335. Madison: The University of Wisconsin Press, 1968.

This evaluation report makes several contributions to the conceptualization of important variables in a cost-benefit analysis of manpower training programs.

A major problem in any cost-benefit analysis is defining the indicators and measurements which will be used. The authors point out, for example, that the perspective of the society and the trainee on the costs of the program may be quite different, yet neither should be ignored. Thus, they use two measures for all of the costs and benefits chosen as indicators. Another problem in measuring benefits is the choice of the major indicator. Most studies have used employment rate after the training program, but the authors argue that wage levels are more mean-

ingful. In measuring the outcomes, fluctuations during the year-in employment levels, general economic activity, etc., were controlled by making comparisons between the trainees and the control group at several points in time.

A third major problem was finding a comparable control group, since true experimental conditions were impossible. The control group selected was drawn from the lists of the unemployed who would have been eligible for the program, but had had no contact. These were selected randomly and later matched into general control categories (sex, age, previous employment record, etc.) with the trainee group. Despite the care taken to match the two groups, the trainees were clearly the "cream of the unemployed." Another point of incomparability lies in the fact that special efforts were made to place a number of the trainees after the program. Also, under conditions of low employment it is possible that trainees took jobs away from non-trainees, and thus artificially lowered the employment rate of the control group. After examining all of these factors, the authors conclude that the control and trainee group were indeed comparable enough to make general comparisons, although not to draw precise conclusions about degree of impact.

In the cost-benefit analysis the authors show that the returns from the program amortize the costs of training within a year. The "rate of return" and the net expected capital value were computed, and values were large. The authors conclude that the size of the benefit-cost ratios to some extent reduces the potential error associated with the non-comparable control group. Another finding which emerges very clearly is that there are considerable differences in the measured benefits of the training program for the several socio-demographic groups. Women and the more well educated showed far fewer gains than males and those with less than 12 years of education. Age was also related to benefits, but not linearly. An appendix deals with the question of whether retraining *creates* jobs as well as fitting people to fill existing jobs.

Caldwell, Michael S. "An Approach to the Assessment of Educational Planning," *Educational Technology*, vol. 8, No. 19 (1968), 5–12.
This paper differs from others in the Bibliography by focusing on the assessment of the planning process rather than on the evaluation of operating programs. Caldwell discusses the total framework of program development: (1) identification of needs, (2) development of strategies and specific plans, (3) implementation of the selected approaches, and (4) evaluation of outcomes. He stresses the need for assessment of the second stage and identifies key decision points in the planning process. He proposes eight criteria for assessment of program planning: relevance, legality, congruence with value systems, legitimacy, compatibility with agency goals, balance, practicability, and cost/effectiveness. For each criterion he suggests appropriate questions to ask and possible methodological approaches for collecting relevant data.

Campbell, Donald T. "Considering the Case Against Experimental Evaluations of Social Innovations," *Administrative Science Quarterly*, vol. 15, No. 1 (1970), 110–113.
This article is a commentary on the Weiss and Rein article, "The Evaluation of Broad-aim Programs." Weiss and Rein point out weaknesses in would-be experimental program evaluations, including criteria chosen for convenience rather than appropriateness, inappropriate control groups, no-effect outcomes, and neglect of the study of process. They conclude that the experimental approach should be replaced. Campbell agrees that these weaknesses should be avoided, but maintains that the correction of the weaknesses is compatible with and even requires the experimental method.

Campbell, Donald T. "Reforms as Experiments," *American Psychologist*, vol. 24, No. 4 (1969), 409–429.
A major constraint on social experimentation is that reforms tend to be advocated as though they were bound to be successful. The "trapped administrator" has so committed himself to the program that he cannot afford an honest evaluation. An alternative stance, commitment to the solution of problems rather than to specific programs, would help to alleviate the political problems associated with evaluation.

The author presents three quasi-experimental designs that are suitable for evaluation. The *interrupted time series design* can be used where data are obtainable for only one case, but where they can be collected for a period of time both before and after the program. The design is weak and is useful only in discerning rather drastic changes.

A *control series design* can be used when time series data are available for two or more cases, only one of which receives the treatment. This design controls for the effects of history, maturation, and test-retest effects. Although it is considerably stronger than the time series alone, data must be obtained over a long enough period to control for maturation-selection effects.

The *regression-discontinuity design* is useful where it is impossible to randomize participants into experimental and control groups. This rarely used design is explained in detail and its potentials are discussed.

Although designs such as these are acceptable under difficult conditions, the true experiment is preferable. The author suggests that conditions necessary for experimentation can be obtained through staged innovation, where those who receive the program later become controls for the early receivers, and through wiser use of pilot projects.

Campbell, Donald T., and Albert Erlebacher, "How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful," in *Compensatory Education: A National Debate*, vol. III of *The Disadvantaged Child*, edited by J. Hellmuth. New York: Brunner and Mazel (forthcoming 1970).
The authors argue that the reason that most evaluations of compensatory education show no effects, or even negative effects, is that they are biased in their statistical analysis. There are several reasons for this, which are discussed in detail with examples. Control groups are not selected by randomization, but by matching with individuals from a superior population. When this occurs, it is difficult to find a statistical technique which will not involve regression of the matched group (chosen on the basis of the extremity of their scores) to the mean. That is, strictly by chance, post-test scores of the matched controls will be higher than their pre-tests, and thus make the change in the "experimentals" look poor. This is particularly true of ex-post-facto studies (most of the compensatory education evaluations fall into this class) and of quasi-experimental studies where pre-tests are not similar in composition to post-tests. Where the tests are similar, common factor covariance adjustments which have been recently developed may be appropriate. The biases in these studies result from poor design, and are thus the fault of the social scientist. The authors encourage more experimentation as a means of avoiding flawed evaluations.

Campbell, Donald T. and Donald W. Fiske, "Convergent and Discriminant Validation by the Multitrait-Multi Method Matrix," *Psychological Bulletin*, vol. 56, No. 2 (1959), 81–105.
This relatively technical paper presents a process for validating tests. It utilizes a matrix of intercorrelations among tests that represent at least two traits, each measured by at least two methods.

Measures of the same trait should correlate higher with each other than they do with measures of different traits involving separate methods. Ideally, these validity values should also be higher than the correlations among different traits measured by the same methods. The notions of convergence between measures of different traits are compared with previously published formulations, such as construct validity and convergent operationalism. Problems in the application of this validational process are considered.

Campbell, Donald T. and H. Laurence Ross, "The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis," *Law and Society Review*, vol. 3, No. 1 (1968), 33–53.
This paper is an evaluation of a program to reduce traffic deaths through stringent enforcement of speeding laws. The evaluation involved gathering official statistics on deaths and accidents over the period from five years before the crackdown through four years after, for both Connecticut and surrounding control states. An extensive discussion of potential interferences with the validity of such statistics is given; in this case regression and instability effects might be operating. The crackdown occurred in a year with a disproportionately high number of deaths relative to former

trends, and a drop might therefore have been predicted even without the crackdown; and there were fluctuations in the annual data over time.

A comparison with surrounding states, however, reduces the cogency of the regression argument since the difference between Connecticut and other states widens after the crackdown. Problems of contamination of data in adjoining states are discussed, as is the difficulty of finding significance tests suitable to these types of data.

The authors conclude that although such a quasi-experimental design does not allow the stringency of evaluation found in the true experiment, rigorous examination of the data can test plausible hypotheses and examine unanticipated consequences.

Campbell, Donald T. and Julian C. Stanley, "Experimental and Quasi-Experimental Designs for Research on Teaching," *Handbook Of Research on Teaching*, edited by N. L. Gage, pp. 171–246. Chicago: Rand McNally, 1963. Also reprinted as *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally, 1966.

This classic paper catalogues three pre-experimental, three experimental, and ten quasi-experimental research designs, and analyzes how well each design protects against the effects of extraneous variables on outcome measures. Eight threats to internal validity are listed: (1) history—outside events which may occur between the before and after measures; (2) maturation—processes within the respondents that occur as a natural consequence of the passage of time; (3) testing—the effects of taking a test once upon the respondents' later scores; (4) instrumentation—changes in the calibration of a measurement device; (5) statistical regression—the tendency of extreme scores to revert toward the mean; (6) selection—biases resulting from differential recruitment into the experimental and control groups; (7) experimental mortality—different rates of loss of people from the experimental and comparison groups; (8) selection-maturation interaction—heightened effects resulting from the combination of selective recruitment and maturation. Four factors are listed that jeopardize

external validity, or representativeness of the results.

Experimental designs control against all these possible sources of interference, but quasi-experimental designs generally leave one or several uncontrolled. The authors examine each design in detail and note ways in which it can be adapted to serve the research purpose. When using quasi-experimental designs it is necessary to identify plausible alternative hypotheses which might account for the outcomes, and to attempt to rule these out. (For example, see Donald T. Campbell and H. Laurence Ross, "The Connecticut Crackdown on Speeding: Time Series Data in Quasi-Experimental Analysis," Section II.)

Caplan, Eleanor K. "Evaluation Research on the Interorganizational Level." Paper presented at 61st Annual Meeting of the American Sociological Association, August 29–September 1, 1966. (*Unpublished*)

Drawing on her experience in evaluating the coordination of 23 community school welfare agencies participating in a juvenile delinquency prevention program, the author discusses the problems of doing research when the units of analysis are complex organizations. These problems are similar in kind, but greater in degree, from the usual issues in research. The first and most crucial task is delineating the research problem within a frame of reference, followed by problems of design, measurement, and analysis.

Each of these steps is further complicated by the following factors: (1) The unlimited variation in evaluation in action programs. (2) The uncharted area of interorganizational analysis. The complexity and variation of the units under study and the lack of theoretically relevant literature leave the researcher groping. (3) The milieu within which the units interact include the total environment of the community. The relationships among the units are affected both by the range of local, state, and national cultural patterns and by the variety of individual roles within the

agency. The possible forces affecting inter-agency relationships within and outside the units are too complex and numerous to include in their entirety. Since the advancement of social science empirical knowledge about inter-organizational relationships can only be obtained through community studies of this type, several suggestions are made about handling the problem areas.

Caplan, Nathan, "Treatment Intervention and Reciprocal Interaction Effects," *Journal of Social Issues*, vol. 24, No. 1 (1968), 63–88.

This study was designed to measure the process of behavioral change of 109 13 to 18-year-old delinquent boys who were counseled by experienced street gang workers. An adjustment classification scale was used, with chief interest focused on the later stages which represented a greater degree of personal adaptation and commitment to program objectives. The stages considered were (5) receptivity to personal counseling, (6) reaching a meaningful relationship, (7) commitment and preparation for change, and (8) transfer and independent manifestation of change. Another scale measured the level of input of the worker from 1 (minimal input) to 6 (supreme effort).

Subjects were included in the study after they had been classified at stage 5, and they were studied periodically for 12 months. One year after having reached stage 5, 6% were at stage 8, 34% at stage 7, 39% at stage 6, 10% at stage 5, and 12% stage 4 or lower. The small proportion of stage 8 success cases is not due to the arrest of a positive change pattern. Post-stage 7 classification changes occur but most are not in the anticipated direction: the trend is predominantly negative, followed by rebounding and successive backsliding. Thus, data taken at any one point will include both climbers and backsliders in each category, nullifying statistical interpretations which assume that subjects at each classification level are similar.

The correlation of positive adjustment and program input for stages 5 and 6 is positive. At stage 7, however, increased worker input is slightly more likely to lead to backsliding. Negative client change is generally followed by increased input by the counselor, followed by increased negative change, followed by greater increased input, etc. The major finding is thus:

over time subjects repeatedly demonstrate a tendency to nearly succeed in adopting change behaviors advocated by the program. They repeatedly fail when faced with the test of experience. Several explanations are offered to account for these results, among them goal incongruity, worker behavior, subject behavior, and interaction effects which result in non-adjustive behavior on the part of both boy and worker.

Caro, Francis G. "Approaches to Evaluative Research: A Review", *Human Organization*, vol. 28, No. 2 (1969), 87–99.

This paper summarizes major themes that have appeared in recent social science literature about evaluation research. *Basic issues* include the definition of the role of the social scientist in planned change and action research and conditions which facilitate the development of research-action relationships. Six obstacles to effective relationships between the research and action spheres are listed: (1) The service orientation and the research orientation often conflict. (2) The time perspectives of the researcher and the practitioner differ. The former tends to look for long-range trends, while the latter is more interested in short-range effects. (3) Practitioners tend to use less rigorously objective or quantitative methods of evaluation and may be unimpressed by the researchers' stress upon them. (4) Researchers have an interest in encouraging change, while practitioners have an interest in the status quo. (5) Researchers are more likely to question the theoretical premises of programs in explaining effects, while practitioners see the problems in terms of lack of resources or individual failures. (6) The basis of experience between the two is different, which may lead to lack of understanding.

Methodological Issues include: identification and measurement of dependent variables, "contamination" of the design, adequate control, sample size, and a general tendency to measure outcomes rather than processes. The logic of the design of evaluation studies is simple, but the methodological problems are enormous. Compromises must be worked out between imperfect situational characteristics and laboratory-type rigor. *Administration of Research* poses many problems. The general consensus in

the literature is that outside evaluators will achieve greater objectivity than staff members of the program, but outsiders run the risk of imposing inappropriate research designs and face an inability to secure cooperation. Their neutrality may be a problem, too, because his "lack of commitment" may exacerbate areas of conflict. Availability of funds is also an issue, since action organizations usually operate under tight budgets. *In implementing findings*, it is noted that the researcher might align himself with people in positions of authority to assure use of study results. Often, the researcher must build up his image as a competent scientist in order to gain acceptance of his research. Questions of interest to the practitioner must be addressed, and early appearance of the results will aid in the contribution to decisionmaking. But the advantages of early feedback must be weighed against the need to control the environment. Attractive presentation may help, as may involving the administrators in the evaluation. The author concludes that the researcher must sensitize himself to the problems he is likely to face if evaluation is to be useful for scientific and practical purposes.

Clark, Burton R. *The Open Door College: A Case Study*. New York: McGraw-Hill Book Co., 1960.

This book is an intensive case study of the development of San Jose Junior College. It is an application of methods of institutional analysis to an educational organization to analyze patterns of organization and interaction with internal and external environments and, more importantly, to assess the impact of the organization on its students. The public junior college was studied because a number of alternative orientations are possible for an institution which is legally a public school but educationally higher education.

The first chapter discusses the administrative setting of San Jose Junior College and its problems. The college had difficulty in establishing itself in a web of school district administrations which were unfavorable to it. Chapter 2 concentrates on the student clientele as it is shaped by admission policy. The wishes and needs of the students affected the purpose of the college. A basic problem of the junior college is that there are students with intentions to transfer

and those with terminal goals; the college has difficulty dealing with both, as witnessed by the small number of graduates. Chapter 3 traces the formal organizational structure, the composition and orientation of the administrative staff, and the building of a faculty. In Chapter 4 the previous data are used to interpret the organizational character of the junior college as a "mass enterprise" defined by its dependency on a large, nonselected, voluntary clientele. It is seen to be heavily oriented toward a secondary school model of organization. Chapter 5 describes the role of the junior college in higher education. The problem is: How can an educational institution be both a public school and a college? The three aspects of status, identity, and autonomy are discussed.

The research was done by unstructured interview, observation, and secondary sources. Records, documents, and memoranda were the best sources of dependable material.

Cochran, William G. *Sampling Techniques*. New York: John Wiley and Sons, 1953.

The emphasis of this book is on sampling theory rather than practice. Topics covered include: simple random sampling; sampling for proportions and percentages; estimation of sample size; stratified random sampling; ratio estimates; regression estimates; systematic sampling; type of sampling unit; subsampling; double sampling; and sources of error in surveys. An assumption is made in the text that the reader has a good grasp of elementary statistics and calculus.

Coleman, James S., Ernest Q. Campbell, et al. *Equality of Educational Opportunity*. Washington, D.C.: Government Printing Office, 1966.

The purpose of this evaluation was to determine the extent to which equal educational opportunities are available to all American children, regardless of their race, and whether differences in education achievement may be traced to differences in the school environment. The study design was non-experimental and ex post facto: aptitude and achievement tests were administered to 645,000 pupils and survey questionnaires were sent to public school teachers, principals, district school superintendents, and pupils in 4,000 public schools. The sample

of schools was stratified by major regions of the country, and along the urban-nonurban dichotomy. Multiple regression analysis was used to determine the relationship of dependent variables, such as the quality of teachers, to the independent variable of achievement. Aspects of the study of particular interest to the evaluator include the sheer magnitude of the study, the measures used, and the sophistication of the analysis.

Results indicated that while school characteristics have a modest effect on student performance, much more of the variation is accounted for by student characteristics and attitudes (such as family background, feelings of efficacy) and aggregate student body characteristics (such as percentage of students who are white). Also of interest is the subsidiary finding that aptitude, or I.Q. tests show more variation among schools that do "achievement tests," and are more sensitive to the effects of school characteristics.

Colvin, Charles R. "A Reading Program That Failed—Or Did It?" *Journal of Reading*, vol. 12, No. 2 (1968), 142–146.

This paper is an account of a study which tried to avoid contamination of subjects by the Hawthorne effect, but failed. The study concerned (1) whether a relationship existed between grade-point average (GPA) and a reading and study skills course for selected college freshmen, (2) whether college life in general leads to improved reading skills for selected freshmen, (3) whether freshmen who take a reading course make significantly larger gains than those who don't.

Findings: (1) There is no significant difference in mean GPA's of freshmen who did take the course and those who did not. (2) Significant gains were shown by all freshmen after one semester, but it is obvious that "college life in general" did not bring about the improvement. (3) Although freshmen in the experimental group improved more on the retest than the controls, the improvement was not significant. It was later found that the control group had been seriously contaminated through advice from advisers, teachers, and help from students in the experimental-group. Such contamination poisons this type of research and demonstrates the need to re-do the study with tighter con-

trols. Too many researchers fail to realize that the Hawthorne effect can operate on a control group as well as on an experimental group.

Community Council of Greater New York, Research Department, *Issues in Community Action Research*, Report of the Spring Research Forum on Evaluation Efforts in Three New York City Community Action Programs, 1967.

Three papers read at the conference are presented: James Jones' report on the status of evaluation at Haryou-Act, Joseph Bensman's paper on the Bedford-Stuyvesant Youth in Action evaluation, which discusses the perils of attempting evaluation when the program is in flux, and Richard Cloward's discussion of the Mobilization for Youth evaluation, which focuses on the difficulties of securing acceptance and use of evaluation findings. Also included are the introductory and concluding remarks of the chairman, S.M. Miller, the group discussion, a list of related readings, and a paper by Terence Hopkins on evaluation strategy.

Cox, David R. *Planning of Experiments*. New York: John Wiley and Sons, 1958.

This book is about the planning of experiments in which the effects under investigation tend to be masked by fluctuations outside the experimenter's control. Although it is oriented primarily toward technological and biological experiments, some of the simpler methods described would be useful for social research. Statistical and mathematical technicalities are avoided in the text, and numerous examples are given to illustrate the principles under discussion. Of particular interest is the fact that several topics, which are not often well covered in the elementary literature on experimental design, are given a good deal of attention. These include such things as the justification and practical difficulties of randomization, the relation of covariance to randomized blocks and to the calculation of adjustments, the different kinds of factors that can occur in factorial experiments, the choice of the size of experiment, and the different purposes for which observations may be made. Chapter headings are: Key Assumptions, Designs for the Reduction of Error; Use of Supplementary Observations to Reduce Error; Randomization; Basic

Ideas about Factorial Experiments; Design of Simple Factorial Experiments; Choice of Number of Observations; Choice of Units, Treatments, and Observations; More About Latin Squares; Incomplete Nonfactorial Designs; Factional Replications and Confounding; Crossover Designs; Some Special Problems.

Cronbach, Lee J. "Evaluation for Course Improvement," *Teachers College Record*, vol. 64, No. 8 (1963), 672–683. Also reprinted in *Readings in Measurement and Evaluation*, edited by Norman Gronlund, pp. 37–52. New York: Macmillan, 1968.

Three functions of evaluation in education are identified: (1) Course improvement—deciding what materials are satisfactory and what improvements need to be made; (2) decisions about individuals—judging pupil merit for selection purposes; and (3) administrative regulation—judging how good the school system is. In the latter two cases, traditional testing procedures are adequate, but for course evaluation it is necessary to locate specific areas where revision is desirable. Total test scores have limited utility. Item data within each test will yield more information about specific aspects of course effect. Tests which are course-specific are limited in their usefulness, since they cannot be used on control groups of students who have not taken the course. Tests should also be developed which measure ability to handle concepts, rather than mere knowledge of factual data. Course evaluation should contribute to the general body of knowledge about learning so that principles of course development may be generated.

Evaluation is too often equated with the administration of formal tests. There are many other methods for examining pupil performance, and pupil performance is not the only basis for appraising a course. Other approaches to evaluation include process studies, which are concerned with events taking place in the classroom; proficiency and attitude measures, which document changes in the pupils; and follow-up studies, which pursue the later careers of those participating in the course.

Cumming, Elaine and John Cumming, *Closed Ranks: Study of Mental Health Education*,

Cambridge, Mass.: Harvard University Press, 1957.

A community educational program was designed to (1) diminish feelings of distance and estrangement from former mental patients, and (2) increase feelings of responsibility for the problem of mental illness. Mental health educational material was presented in small group discussions and via mass media in the form of talks, films, and radio programs, on subjects from infant care to the nature of functional psychoses. The three principles emphasized were (1) behavior is caused and understandable, (2) there is a continuum between normality and abnormality, (3) there is a wider variety of normal behavior than is generally realized.

At the conclusion of six months, the program had reached about 56% of the adult population in the experimental community in western Canada, population about 1,500. A control community was selected from the same province. Initial and follow-up questionnaires measuring attitudes toward mental illness were administered to the entire adult population of the experimental community and to a random sample of the control population. The questionnaire contained two attitude scales, measuring social distance and social responsibility toward the mentally ill. Before and after interviews were held with random sample of adults in the experimental community.

The experimental community responded to the program with anxiety, manifested first as apathy, then withdrawal, and finally hostility. The average community score on the two scales did not change and was not distinguishable from the control community. It was concluded that the six-month program had produced virtually no change in the general attitudes of the population toward the problem of mental illness or toward the mentally ill.

Davis, James A. *Great Books and Small Groups*, Glencoe, Illinois: The Free Press of Glencoe, 1961.

This book presents the evaluation done by the author of the Fund for Adult Education's Great Books program. The evaluation investigated why some of the groups were successful and self-sustaining over a number of years, while others were short-lived. The design of the re-

search involved sampling 172 total groups (1909 individuals), rather than sampling individual participants in the program. Information was gathered through self-administered questionnaires. The focus of the analysis was on the effect of the composition and role structure of the group upon continued participation by members. Variables considered include personal and social attributes of the participants (age, education, involvement in community affairs, etc.), as well as group structure.

Davis, James A. "Great Books and Small Groups: An Informal History of a National Survey," in *Sociologists at Work: Essays on the Craft of Social Research*, edited by Phillip E. Hammond, pp. 212–234. New York: Basic Books, 1964.

The nature of the client-researcher relationship affects the course of evaluation studies. Usually the interests of the two are opposed: the client has *specific questions* which he wishes answered, while the researcher is more concerned with posing interesting *theoretical questions* which will provide him with material for scholarly publications. Compromise and negotiation go on between the parties. The author shows how early discussions of this type affected the texture and character of his well-known evaluation of the Great Books discussion groups.

Evaluations are often serendipitous in nature. Many of the most interesting findings emerge as a result of decisions which are expedient rather than ideal. In this case, for example, the research team wished to take a national sample of all participants in Great Books discussion groups, but there was no listing of individual members. The sample thus consisted of groups, which yielded much interesting information about the organizational characteristics associated with success and failure of this type of voluntary association.

Dexter, Lewis A. "Impressions about Utility and Wastefulness in Applied Social Science Studies," *American Behavioral Scientist*, vol. 9, No. 6 (1966), 9–10.

Dexter identifies the difficulties in doing problem oriented research: (1) The purposes for which the research is contracted may be non-scientific: to postpone a decision which is un-

comfortable to make; to put the onus of a decision which has already been reached on someone outside the agency; to provide officials with the satisfaction of supervising and using scholars, or to provide "scholarly" support for a policy which has already been decided on. (2) Research is very likely to be wasteful, because basic questions are often not formulated. Thus, the results may not be pertinent to the real source of concern. (3) In some cases, the "question" which is to be answered is a sensitive one. The researcher himself must determine what it actually is, without explicitly stating it. (4) The social scientist may offend or upset a good many people, because he probes into sensitive areas. This negates his effectiveness; although, if he has a successor, the knowledge which he has gained may be utilized. (5) Most research is contracted too late to be really useful in policy formulations. (6) Research will not help administrators in the final choice between two unpopular alternatives. (7) In determining the usefulness of a piece of research, it may be helpful to hypothesize the effect of all conceivable results within the utilizing agency.

DiLorenzo, Louis and Ruth Salter, "An Evaluative Study of Prekindergarten Programs for Educationally Disadvantaged Children: Follow up and Replication," *Exceptional Children*, vol. 35, No. 2 (1968), 111–120.

This report summarizes findings after two years of a four-year longitudinal study in New York State of the effectiveness of programs for disadvantaged preschoolers. 1,235 children were enrolled in eight districts, each of which had its own curriculum. The children attended the programs for a year, 2½ hours a day. To test effectiveness, the Stanford-Binet, Peabody Picture Vocabulary Tests, and Illinois Test of Psycholinguistic Abilities, were administered at the beginning and end of the year. Metropolitan Readiness Tests were administered in the late spring of the kindergarten year. Analysis was made in terms of socioeconomic status, district, sex, and race. The prekindergarten experience was beneficial as indicated by differences between experimental and control districts on the Stanford-Binet, PPVT, and ITPA. The most effective programs were those with the most specific and structured cognitive activities.

Whites profited more than nonwhites, although nonwhites benefited significantly. The kindergarten experience sustained the benefits of the prekindergarten year but did not build upon them. However, nonwhite children did not maintain the advantage over their controls and were significantly different from white experimentals at the end of kindergarten. It concluded that more attention should be given to prekindergarten programs, and special programming for the disadvantaged must be carried through the early grades to have lasting value.

Donabedian, Avedis, "Evaluating the Quality of Medical Care," *Milbank Memorial Fund Quarterly*, vol. 44, No. 3 (1966), part 2, 166–203.

This paper describes and evaluates methods for assessing the quality of medical care. It deals with the medical care process at the level of physician-patient interaction.

The outcome of medical care (recovery, survival, etc.) is often used as an indicator of the quality of care. It is easy to measure, but may not be relevant in many cases. The process of care is sometimes used as the approach to assessment and asks the question "Is medicine being properly practiced?" The assessment of settings and structure in which care is given is another possible approach, but the relationship between structure and process, or structure and outcome, is unknown.

Four methods are especially useful for collecting data for measuring care—clinical records, direct observation by qualified colleagues, behaviors and opinions, and reputational surveys. Studies of quality are often concerned with care provided by a specific group of providers, actual care received by a group of people, or the capacity of a group of providers to provide care. Both empirical and normative standards are used to measure quality. Reliability and bias are important problems in interpreting results.

The ultimate test of validity is the effectiveness of the care or the outcome in terms of health and satisfaction. The search continues for easily measurable, reasonably valid indices of medical care, but most existing indices are limited. Suggestions are made for further study —e.g., more conceptual and empirical exploration of the definition of quality, study of the

influences of bias, evaluation of current processes of evaluation. The conclusion is that emphasis must be shifted from evaluating quality to understanding the medical care process itself.

Dorfman, Robert, "Introduction" in *Measuring Benefits of Government Investments*, Washington, D.C.: Brookings Institution, 1965, 1–11.

This section is designed to introduce a series of papers on problems of appraising the benefits likely to accrue from public projects. The benefit-cost framework of project analysis provides the background for the rest of the book and, thus, is described in the introduction.

Government initiative is called for when private enterprise deems unprofitable an investment which is socially worthwhile. This is especially true when a facility or service is a collective good or the act of consumption of the facility or service is a collective good. Other incentives, like preservation of natural resources, or desire to reduce inequalities, may also lead to government action. But inherent in government enterprises is the fact that market prices cannot be used in appraising their contributions. Thus the need for benefit-cost analysis.

The starting point of such an analysis is a projection of the physical output of the undertaking in a given year of operation. One approach is to calculate the gross benefit for that year and compute the ratio of gross benefits to total costs. An alternative is to subtract current costs in the year from gross benefits to obtain current net benefits. The net benefits for each year are added up. This figure is then compared with the estimated capital cost to obtain the benefit-cost ratio. Debate centers on the question of whether social benefits can be estimated reliably enough to justify the trouble involved in a benefit-cost computation.

Downs, Anthony, "Some Thoughts on Giving People Economic Advice," *American Behavioral Scientist*, vol. 9, No. 1 (1965), pp. 30–32.

Most clients do not understand the elementary economics of information. The logical purpose of advance research is to keep from making expensive mistakes. Too often the value of doing research is underestimated by the

client and overestimated by consultants who become fascinated with the problem and do more research than is economically justified. When research is undertaken, frequently clients need redefinition of the problem and the suggestion of alternative approaches. Some clients have an exaggerated idea of the precision with which economic advice can be given. Consultants also run into problems when people seek professional advice not because they need help in decision-making, but to settle internal disputes, justify conclusions already reached, discredit a rival or entrench a position, or as an excuse for not acting at all. Consultants must be aware of both secondary and purely technical issues at stake in order to give the best possible advice.

Dressel, Paul L. *Evaluation in General Education*, Dubuque, Iowa: Wm. C. Brown Co., 1954.

This book is a collection of evaluation reports on general education. General education is defined as a set curriculum of classics and liberal arts courses on the college level. Goals for these programs, although somewhat diffuse and varied in the different colleges and universities covered in the book, include such items as development of critical thinking, understanding other people, providing the basis for wise vocational choice, development of self-understanding, etc. Measurement techniques and designs of evaluative studies also vary widely among the different schools.

Dressel, Paul L. (Ed.), *Evaluation in Higher Education*. Boston: Houghton Mifflin Company, 1961.

Eleven authors treat aspects of evaluation relating to college and university educational programs, including the nature and objectives of evaluation, procedures for evaluation particularly of departmental programs, and institutional self-evaluation.

Drew, Elizabeth B. "HEW Grapples with PPBS," *The Public Interest*, No. 8 (Summer, 1967), 9–29.

This article discusses the history of the Planning-Programming-Budgeting System (PPBS) in the department of HEW and the results of

the initial studies they have done. During 1966–67 HEW completed four PPB studies: (1) of selected disease control programs (automobile accidents, cancer, arthritis, syphilis, and tuberculosis), (2) of human investment programs in rehabilitation and training, (3) of maternal and child health programs, (4) of potentials for improving income maintenance.

All the studies were severely limited by a lack of basic information and statistics. Analysis of the disease control program was relatively helpful. Using the criteria of cost per death averted and benefit-cost ratio, the study measured the impact of all the disease control programs. From the analysis, a priority ranking for the use of funds was drawn up. In the other studies, such categorization was impossible; there was even less information available. In the human investment programs, benefits were defined as the increase in earnings which resulted from each of the programs, but benefits were difficult to pinpoint. It was concluded that vocational rehabilitation programs deserved to be expanded, but all human investment programs must be improved. The analysis of proposed programs in child health and income maintenance were even more difficult to conceptualize and study.

The PPB system is still new and quite crude but it is a "giant stride forward" judged against traditional bureaucratic decision-making.

Dyer, Henry S. "The Pennsylvania Plan: Evaluating the Quality of Educational Programs," *Science Education*, vol. 50, No. 3 (1966), 242–248.

This condensation of a three-volume report to the Pennsylvania State Board of Education discusses the purposes and elements of a plan to evaluate the school program's capacity to meet the needs of the children. The plan included five steps: (1) Ten major goals of education were defined. (2) Measures were located and more developed on detailed goals of instruction. These were primitive but useful. (3) Performance standards were devised, using the measures available. The average of the output of the schools at the top of the range in any given category served as the standard for all schools in the category. (4) An educational research program was incorporated into the ongoing evaluation program. Both research and

evaluation occur continuously. (5) When results begin to come in, they will be used to strengthen each school district's educational program.

Eaton, Joseph W. "Symbolic and Substantive Evaluation Research," *Administrative Science Quarterly*, vol. 6, No. 4 (1962), 421–442.

Ambivalent attitudes toward evaluative research often arise in large organizations because evaluation involves spending money which could be used for more immediate purposes, and it may produce disturbing data on organizational problems. Research has two very different functions for organizations: symbolic and substantive. The latter can occur when there is a scientific interest in asking questions, gathering data, and interpreting and communicating the results. The former results from the conflict between two attitudes—the belief in the value of exploring the unknown, and the fear of disturbing positions of power or raising questions about existing agency operations. The research is done, but its results are likely to remain uninterpreted and uncommunicated. This hypothesis is supported by data from a social work organization which shows that even where personnel evince interest in conducting research, there is little willingness to interpret research findings if they appear discouraging. Verbal communication channels are preferred over written, and among written channels, internal publication is preferred to public release of findings. Self-censorship seems to occur at all levels, despite the official pro-research policies of the organization and administrative dependence on information. Bureaucratic impediments —loyalty to the organization and fear of consequences of negative findings—appear to account for the high incidence of symbolic research.

Educational Evaluation: New Roles, New Means; The 68th Yearbook of the National Society for the Study of Education, R. W. Tyler, ed. Chicago: NSSE, 1969.

This is a collection of papers on the subject of educational evaluation. Titles and authors include:
"Historical Review of Changing Concepts of Evaluation," Jack C. Merwin
"Some Theoretical Issues Relating to Educational Evaluation," Benjamin S. Bloom.

"The Uses of Evaluation in Guidance," Ralph F. Berdie
"Research in College Admissions," Dean J. Whitla
"Evaluation and the Award of Scholarships," John M. Stainaker
"The Evaluation of Group Instruction," Herbert A. Thelen
"The Role of Evaluation in Programs for Individualized Instruction," C. M. Lindvall and Richard C. Cox
"The Relationships between Research and Evaluation Studies," John K. Hemphill (see annotation under Hemphill)
"The Uses of Educational Evaluation in the Development of Programs, Courses, Instructional Materials and Equipment, Instructional and Learning Procedures, and Administrative Arrangements," John C. Flanagan
"Evaluation of Ongoing Programs in the Public School System," Malcolm Provus (see annotation under Provus)
"Appraising the Effects of Innovations in Local Schools," Henry M. Brickell
"Evaluation in Assessing the Progress of Education to Provide Bases of Public Understanding and Public Policy," Jack C. Merwin and Frank B. Womer
"International Impact of Evaluations," Torsten Husén
"The Impact of Machines on Educational Measurement," E. F. Lindquist
"Needed Concepts and Techniques for Utilizing More Fully the Potential of Evaluations," Robert E. Stake and Terry Denny
"Outlook for the Future," Ralph W. Tyler

Educational Testing Service, *Disadvantaged Children and Their First School Experiences: ETS-Head Start Longitudinal Study*. Princeton, New Jersey: Educational Testing Service, August 1969.

This report provides a history of the study from 1967 to summer 1969. Descriptive reports from each of the four communities included in the study are given. Separate chapters discuss measures to be used (including background of children, interaction of children and mothers, tests of cognitive and personal-social characteristics, medical data, etc.); data collection pro-

cedures; plans for data analysis; plans for 1969–70.

Appendices provide supplementary material on the measures, working papers, description of project personnel, and a short discussion of the Westinghouse Study of Head Start.

Educational Testing Service, *Disadvantaged Children and Their First School Experiences: EST-Head Start Longitudinal Study*. Princeton, New Jersey: Educational Testing Service, February 1970. (*Unpublished*)

This report describes the sample of disadvantaged children included in the study in three cities, and reports on the progress of data collection. Measures include classroom observation ratings, records of the child's school experience, three batteries of child tests, parent interviews, teacher and school administrator questionnaires.

Educational Testing Service, *On Evaluating Title I Programs*. Princeton, New Jersey: Educational Testing Service, 1966.

The abridged proceedings of a workshop on evaluation, attended by 39 participants from 24 states, are presented. The papers and discussion focus on four topics: (1) Title I educational objectives and the role of evaluation, (2) selecting and developing evaluation instruments, (3) designing and interpreting the results of evaluation studies, and (4) problems in evaluation research and suggested solutions.

Elinson, Jack, "Effectiveness of Social Action Programs in Health and Welfare" in *Assessing the Effectiveness of Child Health Services*, Report of the 56th Ross Conference on Pediatric Research, pp. 77–88. Columbus, Ohio: Ross Laboratories, October 1967.

This paper examines 10 published evaluation studies that used control groups in an approximation of classical experimental design and were competently conducted. The significant finding is that none of the 10 studies found striking positive effects. There are a few positive glimmers, but by and large good studies come up with negative results.

The discussion that followed Elinson's paper at the conference is presented. Questions were raised about the validity of the measures used in evaluation research, the need to study what

actually occurs in the course of program interaction, the utility of control-group experimental research for assessing effects of programs, the difficulty of looking at only one type of change in a complex social fabric, the level of expectations for what programs can accomplish.

Etzioni, Amitai and Edward W. Lehman, "Some Dangers in 'Valid' Social Measurement," *The Annals of the American Academy of Political and Social Science*, vol. 373 (September 1967), 1–15.

This paper is a preliminary statement on the dysfunctions that social measurement may have for societal planning. Two broad classes of dysfunctions are identified: (1) arriving at invalid conclusions which become the basis for erroneous policy decisions and (2) ignoring those dimensions and indicators of a concept that are most susceptible to social manipulation. Of particular interest to the evaluator is the discussion of goal models versus system models as bases for evaluation. The goal model, which measures success solely in terms of goal achievement, is deficient because most organizations never fully achieve their goals, and the investigator may be sidetracked from the pursuit of more relevant information. The systems model, on the other hand, is promised on the idea that any organization or program must deal with many recurrent problems besides those directly associated with goal achievement, and that failure or success may be due to a large extent to the allocation of scarce resources among all activities. The systems model is, of course, more complex to use as a basis for evaluation, since it requires detailed knowledge about the organization and its environment as well as its goals, and an understanding of the optimal allocation of resources.

"Evaluating Educational Programs: A Symposium," *The Urban Review*, vol. 3, No. 4 (1969), 4–22.

The symposium, which focuses on Title I of the Elementary and Secondary Education Act (ESEA) and its requirement for evaluation of programs, consists of ten brief contributions by eleven authors.

J. Wayne Wrightstone discusses the problems encountered by the New York City Board of Education in evaluating Title I programs for

the education of disadvantaged children. He suggests the need for external evaluation on sensitive problems and the convenience of internal evaluation on routine problems. All decisions and conclusions reached by the evaluator should be a result of frequent communication between evaluator and program staff. The evaluator is bound to the minimum "givens" of a project, and must accept the mandated objectives determined by the Board of Education.

James S. Coleman discusses the focus of educational evaluation research. The traditional emphasis has been on input—class size, per pupil expenditure, etc. However, such a focus does not reveal what is actually happening. Examination of outputs, primarily in the form of academic achievement, is essential. Also, there is often a difference between inputs as offered and inputs as received. The loss of input between its disbursement by authorities and its receipt by the pupil may be a major explanatory variable in analyzing the effectiveness of programs.

David Hawkridge and Albert Chalupsky base their discussion on their experience with a national survey of programs for the education of the disadvantaged. They discuss an ideal evaluation model and the limitations placed on the model in evaluating educational programs. Ideally educators should develop testable hypotheses, design the experimental program, select the subjects, obtain measures of the program's effectiveness, draw conclusions from the data, and modify the program on the basis of what has been learned. However, data are often untrustworthy, an evaluator's task is often determined by his sponsor, the evaluator cannot impose his own beliefs on the program, and funds are often limited. A workable compromise between the ideal model and the current reality of often shoddy evaluations can be reached by concentrating funds on intensive evaluation of fewer selected programs.

Henry S. Dyer discusses the present desparate straits of educational evaluation. Given the amorphous nature of education, educational evaluation is very complex and rather messy. Evaluators must be in on the planning of projects to help provide clear objectives, non-duplication of research, adequate data collection, and built-in experimental design. But the constraints imposed by primitive measures and the

subordination of evaluation to program operation are severe. Evidence suggests that educational systems are ill-equipped with the requisite personnel and expertise to assess the effects of programs on students. Evaluation, both creative and analytical, must be continuous and Title I's requirement for annual evaluation is a step in the right direction.

John Mann makes the point that evaluators, more than other researchers, must frequently compromise the precision of their findings for the sake of quick and frequent results. The better the study, the longer it takes, and consequently the less use it may have for administrators. The sloppier the method, the more likely it is to provide timely, interesting information, even though the information is of doubtful validity. A compromise must be sought to attain both precision and practicality.

Martin Mayer stresses that the program is more important than evaluation. Evaluation is useful for finding out which direction the children are taking as a result of the program and whether their direction makes sense. But evaluators must give priority to what the program is doing, and they must not limit the program by imposing on it confining objectives, tests, or measures.

Edward A. Suchman develops an evaluation model. Evaluation research asks about the kind of change desired, the means by which the change is brought about, and the signs by which the change can be recognized. It is based on the same basic logic as non-evaluative research; any difference is one of purpose and not of method. The focus should be on finding out whether it was really activity "a" which achieved objective "b" and how and why the activity was able to achieve the objective. Evaluations of success must be made in terms of conditional probabilities involving factors which are only disposing or contributory rather than determining. A program must be viewed as part of an ongoing social system. The ideal study tests under field experimental conditions the hypothesis that activity "a" will achieve objective "b" *because* it is able to influence objective "c" which affects the occurrence of process "c" which affects the occurrence of the objective. There are two possible sources of failure—the inability of the program to influence the causal variable and the invalidity of

---

the theory linking the causal variable to the desired objective.

Peter H. Rossi states that the prestige of evaluation research is lower than that of pure research. Possible reasons for the low standing are: evaluation research is done outside the university; the evaluation researcher is often defined as an employee providing services for an employer; the outcomes of evaluation have little impact on programs; reports are diffused to a limited audience; evaluation research is often of low quality as research. Policy makers must make a commitment to evaluation by developing action alternatives for dealing with evaluation results, and by adopting policies guided by feedback about the success of their programs rather than rigidly adhering to an ideology of operation.

Edward Wynne discusses the importance of directing educational evaluation toward the ultimate users, the community. Administrators currently in possession of evaluation results do not present them to the public. Until evaluation is done for the group that has the highest interest in efficiency in education, i.e. the parents, evaluators will restrict their efforts to narrow, artificial frameworks.

Michael Scriven makes a number of specific criticisms of educational evaluations, including inconsistency of data gathering, parochialism, stress on superficialities, and casual acceptance of unreliable claims. He concludes that evaluators are not accepting their full responsibility. Their reports are generally inconclusive when they should reach solid conclusions about the program's worth.

Evans, John W. "Evaluating Social Action Programs," *Social Science Quarterly*, vol. 50, No. 3 (1969), 568–581.

Relatively little is known about the effectiveness of many of the massive social action programs recently initiated by the federal government. Reasons for the lack of empirical evaluations include: poor methodological tools and measuring instruments; the complexity of the program environments, which makes it difficult to determine causal relationships; lack of real support for rigorous evaluation studies by government administrators and program directors; and the conflict between the requirements of the program and those of effective evaluation.

A "master plan" for evaluation has been developed at OEO, which classifies evaluation into three categories: (1) the assessment of overall program impact and effectiveness, (2) the evaluation of the relative effectiveness of different program strategies and alternative techniques, and (3) the evaluation of individual projects, through site visits and other monitoring activities, to assess managerial and operational efficiency. Experience with this plan has resulted in several important generalizations about evaluations of federal programs: (1) evaluation is not a waste of time; (2) the aim of evaluation is not to produce methodologically perfect studies but to improve decision making; (3) much can be done with existing, imperfect research techniques; (4) evaluation should be made a central part of the management process; (5) evaluation teams should be staffed by professionally qualified personnel; (6) an invulnerable source of funds should be made available for evaluation; (7) evaluators should be aware of the controversial nature of their task, so that they are prepared to deal with harassment and hostility from various sources.

Fairweather, George W. *Methods for Experimental Social Innovation,* pp. 24–36, New York: John Wiley and Sons, 1967.

Experimental social innovation studies are unique. They combine features of five methods: descriptive-theoretical, surveys, laboratory, participant-observer, and service. A study is defined as a social innovative experiment when it (1) defines a significant social problem, (2) carries out observations, (3) innovates a new subsystem, (4) designs an experiment to compare it with the traditional subsystem, (5) places the subsystems in the appropriate social context, (6) is longitudinal to allow for evaluation, (7) makes researchers responsible for participants, and (8) is multi-disciplinary.

Unique design characteristics include: (1) The control subsystem is the usual social practice for the social problem; this is appropriate because substantive evidence is needed to justify replacing the traditional with the new. (2) Agreement between administrators and researchers is needed to permit the experiment to be completed. (3) Social innovation is primarily empirical in nature and requires new approaches to measurement and analysis. (4)

Outcome criteria must be selected which are socially acceptable and meaningful for those acquainted with the problem.

The social innovator faces problems including the need to work under field conditions, responsibility to the participants for results, and the frequent reluctance of administrators to implement findings. A model for executing a social innovation is presented.

Fairweather, George W. *Social Psychology in Treating Mental Illness*. New York: John Wiley and Sons, 1964.

The purpose of this study was to investigate the possibility of forming autonomous problem-solving groups of hospitalized patients to provide mutual support and serve as a bridge to the outside world. The small-group treatment program was instituted on one ward of the neuropsychiatric section of the V. A. hospital in Palo Alto, California. Patients in the program were assigned to four task groups, which met two hours daily. These groups were given responsibility for recommending to the staff a course of action for each member's daily living and future plans. The experimental phase within the hospital lasted 27 weeks, with a follow-up of 26 weeks. Patients for the study were mostly schizophrenics although other groups were represented, and were of varying degrees of chronicity. Matched on age, diagnosis, and length of hospitalization, the patients were randomly assigned to one of two treatment programs, small-group and traditional. The experiment consisted of two-by-four analysis of variance, factorial type, with two treatment and four diagnostic-chronicity groups. A total of 111 patients were in the small-groups, and 84 in the traditional treatment program.

Measures to assess treatment outcome included both hospital criteria of improvement and community follow-up behavior. In-hospital measures were physical and social activity indices, attitudes and perceptions, and a sociometric test. The community measures were rehospitalization, employment, friendships, communication, appraisal of illness, drinking behavior, nature of residence, membership in community groups, involvement in leisure-time activities.

The outcomes of the study led to four conclusions: (1) Small-group patients were physically more active and had a greater degree of social participation. (2) Small-group wards reflected a significantly higher degree of cohesiveness. (3) The treatment program had no effect on attitudes toward mental illness. (4) The small-group program reduced hospitalization significantly and resulted in more employment and active involvement. However, approximately 50% of long-term psychotics returned to the hospital within 6 months.

Fellin, Phillip, Tony Tripodi, and Henry J. Meyer (Eds.), *Exemplars of Social Research*. Itasca, Illinois: F. E. Peacock Publishers, 1969.

The objective of this book is to improve the *consumption* of empirical research reports. It offers the reader advice on how to assess study reports and offers a scheme of classification of research. Major categories of research (experimental studies, quantitative-descriptive studies, and exploratory studies) are described and subtypes are identified. A series of "guideline questions" alerts the reader to issues of classification, problem formulation, research design and data collection data analysis and conclusions, and utilization of results. There is an emphasis on critical assessment of research and sophisticated utilization of research to improve social practice.

Brief reports of 21 studies follow, seven in each of the major classification categories. Several of these "exemplary studies" are evaluations, such as Berleman and Steinburn's "The Execution and Evaluation of a Delinquency Prevention" and Meyer, Borgatta, and Jones' "An Experiment in Prevention through Social Work Intervention" (a brief report of the author's study of *Girls at Vocational High*).

Ferman, Louis A. "Some Perspectives on Evaluating Social Welfare Programs," *Annals of the American Academy of Political and Social Science*, vol. 385 (September 1969), 143–156.

An evaluator's job encompasses two dimensions—his logical investigation of the program and his social interaction with the sponsor of the evaluation and the staff of the agency being evaluated. The problems encountered by evaluators lie primarily in the second dimension. The three sets of actors have varying perspectives on the purpose and conduct of the evaluation, arising from conflicting interests and professional values. There is a need for compromise among evaluator, staff, and sponsor. Such factors as changing goals within the program may force the evaluator to redesign his study. The politics of evaluation may involve him in a situation in which he must defend his research against those who see its results as a status threat or as support for an opposing social ideology. In spite of these issues, evaluation design and analysis still use traditional tools of social science research. Difficulties in performing evaluation research arise not from the research techniques used, but from the interpersonal situation in which the evaluator is placed.

Flanagan, John C. "Evaluating Educational Outcomes," *Science Education*, vol. 50, No. 3 (1966), 248–251.

Evaluation of specific components of the educational program is necessary. Five evaluation methods are given: (1) Ask the students to assess the success of the instructional program. (2) Ask recent graduates to evaluate their education. (3) Evaluate all students using measures of a common set of educational objectives. (4) Measure each student's progress toward his own objectives. (5) After each student has completed his education, ask if his education helped him progress toward his goals.

Recent studies are given as examples of each method. All methods are appropriate and useful, although each can be improved by further research. Four kinds of information are needed for valid evaluation: (1) capabilities of student performance, (2) definitions and predictions of successful educational processes, (3) information on effective educational materials, (4) efficient procedures for evaluating students' progress.

Fleck, Andrew C., Jr. "Evaluation as a Logical Process," *Canadian Journal of Public Health*, vol. 52, No. 5 (1961), 185–191.

Evaluation should be a logical process which relates results achieved to costs incurred. Often critical evaluation is not performed because of the inertia of tradition. In the public health field, evaluations are usually carried out by practitioners, whose personal goals and commitment to the program may interfere with the utilization of findings. Program directors also tend to become preoccupied with means and neglect ends. For example, in a recent evaluation of tuberculosis control, information was gathered about both active and inactive cases because both were of clinical interest, although the program goals were to isolate active cases only. The failure to gear public health programs toward ends occurs because specific epidemiological goals are seldom specified. Valid evaluations cannot be made unless the program has (1) a description of the underlying idea or epidemiological theory, (2) a statement of purpose which is universally understood, (3) a description of the materials, devices, personnel, and processes to be used, and (4) a practice of reporting results which are logically related to the rationale behind the program.

Fleck, Andrew C. "Evaluation Research Programs in Public Health Practice." *Annals of the New York Academy of Sciences*, vol. 107, No. 2 (1963), pp. 717–724.

Research in the evaluation of modern public health practice inevitably involves the study of formal organizations. If evaluations are to be successful—i.e. produce deliberate action—intimate knowledge of organizations and their settings is required. This will involve examining not only their formal structures, but also interaction patterns, individual motives, etc. This is particularly true in the case of intra-organizational evaluation, where the antithetical needs for stability and survival must be assessed. The choice of evaluational method will be dependent on the relative value placed by the organization on each of these two conflicting goals. If the organization values short-run stability, for example, they will be unreceptive to any evaluation which is overly disruptive, and will accept ritualistic or operational evaluations. The task of the outside researcher is to create acceptance, on all levels, of the fact that change is necessary for organizational survival.

Fox, David J. *Evaluation of New York City Title I Educational Projects 1966–67: Expansion of the Free Choice Open Enrollment Program*. New York: The Center for Urban Education, 1967. (*Unpublished*)

The 1966–67 evaluation of the Open Enrollment Program in New York City public schools is a follow-up to the 1965–66 study. It collected two kinds of data not obtained in the original study—in-class observation of lessons in "sending" schools and perceptions of "sending" school principals about the program. Comparable data for "receiving" schools were collected through partial replication of the earlier study.

Forty-one elementary and junior high schools were studied. Lesson observation reports, teacher behavior reports, general school reports (all three completed by observers), reading scores, and principals' interviews were analyzed. Results indicate that in elementary schools there was little difference between ratings of teaching or teacher behavior in sending and receiving schools. However, open enrollment children in receiving schools rated higher in social and personal functioning—but not in academic achievement. In junior high schools, receiving schools were superior in all facets studied.

Results of the two studies together indicate that both children and school staff favor the open enrollment program, that open enrollment children benefit in social and personal functioning, that the more academically able children probably move into the program (thus leaving the sending schools with less able pupils), that there is no steady improvement in reading ability among participants, and that receiving schools do not suffer academically from the entry of open enrollment pupils.

Fox, David J. "Issues in Evaluating Programs for Disadvantaged Children," *Urban Review*, vol. 2 (December 1967), 7, 9, 11.

Fox criticizes the state of evaluation on several grounds: (1) The term "disadvantaged" has never been well defined, and is often equated with minority group status. (2) "Programs" are being evaluated even before they have been developed in the field. A distinction should be made between critical evaluation (judgments about the program's worth) and on-going evaluation (oriented toward feedback and development). (3) Evaluators have used irrelevant criteria for evaluation, such as substituting measures of social functioning for criteria of intellectual progress, using criteria which are beyond the scope of the programs (such as

improved self-image and aspirations). The relative importance of criteria for measuring unintended consequences remain ignored. Researchers have often accepted stereotypes and contested assumptions about children, such as the fact that disadvantaged children always have poor self images. (4) The term progress has not been adequately defined. (5) Researchers have been too little involved in the implementation of their findings. Immediate feedback is often essential; given the complexity of the data, the social scientist has a major responsibility in the implementation process. (6) Researchers have not challenged the current orientation of present programs, which are preoccupied with remediating weaknesses rather than building on strengths.

Freeman, Howard E. and Clarence C. Sherwood, "Research in Large-Scale Intervention Programs," *Journal of Social Issues*, vol. 21, No. 1 (1965), 11–28.

The evaluator of large-scale social programs faces a difficult environment. Programs and organizations are constantly changing, and the projects to be evaluated involve multiple stimuli and multiple goals. In order to evaluate a complex program, it is important to develop an impact model that shows the hypothesized cause-and-effect relationships between program principles, procedures, and outcomes. The model should reflect the theoretical concepts on which the program is based, as well as activities specific to the program, so that it can be used in a comparative context. The researcher must participate in all stages of program development and implementation in order to make sure that the model is realistic and that the program does not alter so drastically that the model is rendered meaningless.

Massive programs should be evaluated not only in terms of outcome (efficacy), but also in terms of accountability (whether the target population is a significant one, whether the program is being effectively implemented, etc.) and efficiency. Measurement problems which should be considered are regression effects, shifts in scores as the result of factors outside the program, and exposure to multiple programs. Evaluation should seek to approximate experimental design. Although random assignment to treatment and non-treatment is rarely possible,

allocation to alternative treatments is usually feasible.

Getting, Vlado A. "Part II—Evaluation," *American Journal of Public Health*, vol. 47, No. 4 (1957), 409–413.

The author reviews the literature on evaluation published by the American Public Health Association, and discusses other health evaluations by national and state groups.

Glaser, Edward M. and Hubert S. Coffey, *Utilization of Applicable Research and Demonstration Results*. Los Angeles: Human Interaction Research Institute, n.d.

This booklet reports the findings of a study on how utilization of the results of demonstration projects in vocational rehabilitation can be facilitated. The study found that presenting research reports in attractive, readable form had some impact, but even more conducive to adoption of innovations were attendance at a conference where potential users could discuss the innovation with program operators, and site visits to see it in operation. The published research report had an effect primarily where there was an existing predisposition to innovation, whereas the conference appeared to break down barriers to the spread of innovation. Sending "missionaries" from the innovative site to potential users who had already read the report, attended the conference, and made a site visit did not appear to increase the utilization rate significantly. Outside consultation to management helped organizations to change more rapidly and seemed to stimulate agencies to search for new sources for innovation.

Glennan, Thomas K., Jr. *Evaluating Federal Manpower Programs: Notes and Observations*. Santa Monica, California: The Rand Corporation, September 1969.

This report discusses three areas of evaluation: (1) the conceptual framework for benefit-cost evaluation, (2) problems in the measurement of benefits and costs, and (3) the relationship of program evaluation to the planning process.

Benefit-cost analysis should be concerned less with average benefit-cost ratios than with marginal benefit-cost ratios, since the issue to be addressed is what the effects will be of an

increase or decrease in funding levels. Measurement of benefits and costs must take into account their distribution among different economic and social classes and the social value of benefits to different groups. The quantification of nonmonetary benefits is important.

Control groups are difficult to obtain. The most reliable controls would be a sample of individuals qualified for a program who for some reason did not enroll. Longitudinal studies are valuable for obtaining current, rather than retrospective, data on both benefits and expenses over time. Cross-program comparisons are extremely useful, but their value as natural experiments is limited by problems of multiple causality and lack of theory for relating psychological variables to performance.

Evaluations of total program impact do not provide the kind of information that can aid in program planning. Although government funding bodies are concerned with total impact as a guide to allocating resources among programs, program managers want rich information on details of program operation. In the early stages of a program, studies should look at program components and procedures; in the later stages evaluation of impact is suitable. Evaluation of individual local projects is very expensive and is not likely to be worth the investment.

Glock, Charles Y., et al. *Case Studies in Bringing Behavioral Science into Use*, Studies in the Utilization of Behavioral Science, vol. 1. Stanford: Institute for Communication Research, 1961.

This monograph contains papers by five producers of behavioral science research and three important users of behavioral research. All are based on case experience with applying social science knowledge. The papers are:

Charles G. Glock, "Applied Social Research: Some Conditions Affecting Its Utilization"

Ronald Lippitt, "Two Case Studies of Utilization of the Behavioral Sciences"

John C. Flanagan, "Case Studies on the Utilization of Behavioral Science Research"

Elmo C. Wilson, "The Application of Social Research Findings"

Carroll L. Shartle, "The Occupational Research Program: An Example of Research Utilization"

M. L. Wilson, "The Communication and Utili-

zation of the Results of Agricultural Research by American Farmers: A Case History, 1900–1950"

George W. Croker, "Some Principles Regarding the Utilization of Social Science Research within the Military"

Howard E. Page, "Research Utilization"

Gollin, Albert E. "The Evaluation of Overseas Programs: Applied Research and Its Organizational Context," in *Education and Training for International Living: Concepts*, edited by Robert Campbell, Bert King and John Nagay, Arlington, Va.: Beatty Publishers, 1970 (forthcoming).

The author emphasizes three aspects of evaluation: the types of data to be collected, types of program objectives that serve as goals or standards, and the manner in which the former are related to the latter. The relationship between the data and the specified goals is particularly crucial in overseas programs, where goals are often very difficult to specify. Thus, he concludes that it is essential to note which standards and criteria are being used, since any one of a number of perspectives may be found.

The steps in an overseas evaluation are enumerated. In *specifying goals*, it is necessary to get a clear picture of what they are, and what end states would be considered especially harmful. Since most of the goals of overseas programs are diffuse, it is usually necessary to *choose instruments* which have a number of indicators for each goal. Care should be taken, whatever the design, to include the *collection of baseline data*. In collecting *post treatment information*, two main problems arise. The first is determining *who* should collect the information, since this will affect the reliability of the responses (for example, agency personnel interviewing native community members might get very different responses from those given to a native interviewer); the second is *limiting* the amount of data collected from each individual so that cooperation will not entail too great an annoyance. In *analyzing and reporting the data*, the researcher should avoid being pressured into presentation before he is ready, and his writing should take into consideration the audience of the report (managerial level, policy-making level, etc.).

In discussing the sources of criteria information, the author rejects the use of supervisor ratings and participant appraisal as limited and often biased. Highly useful in the overseas program context, however, are judgments made by people in the host country and by people who do not have a stake in the success or failure of the program. It is often difficult to find clear or specific measures of output at present, since those tasks which are easiest to measure may not be the best indicators of the programs goals.

Past surveys of overseas program have tended to focus on recruitment and selection of personnel, rather than training and service in the field. However none of these stages has been adequately investigated. Another problem with past research is that it has tended to be prematurely policy-oriented. Given the level of ignorance about these programs, it seems advisable first to attempt intensive descriptive analyses of the content and operation of aid programs. This is particularly true because of the complexity of these programs and their goals. The author also suggests that some time be devoted to studies of the role requirements and role expectations for a "change agent" and the changing characteristics of volunteers for such programs as the Peace Corps.

Gorham, William, "Notes of a Practitioner," *The Public Interest*, No. 8 (Summer, 1967), pp. 4–8.

The author, who was in charge of introducing the Planning-Programming-Budgeting System (PPBS) into H.E.W., discusses the problems encountered. Although the objective of PPBS is to use cost-benefit analysis in the decision-making process, most decisions on government spending emerge from a political process. There are problems concerning the availability of data, definition of benefits, measurement of benefits, weighting of benefits to different people, and the inherent heterogeneity of the benefits in fields like health, education, and welfare. Nevertheless, the very process of analysis is valuable because it forces program administrators to think about their objectives and how they can be measured. Thus analysts can upgrade the programs and help identify successful techniques. Even tentative indications from PPBS will be useful.

Griessman, B. Eugene, "An Approach to Eval-

uating Comprehensive Social Projects," *Educational Technology*, vol. 9, No. 2 (1969), 16–19.

This paper provides a concise outline of the various functions which evaluations may serve and the task-steps which are essential in the research design. Special emphasis is laid on the problems associated with the evaluation of large-scale multidimensional programs.

The techniques used in any particular evaluation will be dependent on the functions which it is expected to fulfill. The primary function is usually the answering of pertinent questions about the adequacy, efficiency and success of the program. Other important functions may be (1) legitimization of a program, (2) desire for feedback information of use in decision making, and (3) the discovery of basic information applicable to related subject areas. The relative importance of these secondary functions should help to determine, in part, the design and nature of the evaluation.

Seven steps in evaluation research are located: (1) problem identification; (2) development of an evaluation model; (3) operational definition of goals; (4) devising appropriate research techniques; (5) collecting the data; (6) analyzing the data; and (7) reporting the findings. Particular attention is paid to steps 2 and 5, where a new evaluation model CIIP is discussed. (See Stufflebeam, Guba, Alkin.) This model is based on four generalized areas of evaluation—context (environment), input (resources), process (what the program did) and product (intended changes in individuals, social relationships, social system balance, and unintended changes). This framework can be used as a basis for both building a specific model of interactional effects of the variables under review and for gathering appropriate kinds of information.

Guba, Egon G. "Development, Diffusion and Evaluation," in *Knowledge Production and Utilization in Educational Administration*, edited by Terry L. Eidell and Joanne M. Kitchel, pp. 37–63. Eugene, Oregon: University Council for Educational Administration and Center for the Advanced Study of Educational Administration, University of Oregon, 1968.

This paper deals with the large gap between knowledge production and utilization and the problems involved in bridging the gap. A theory-practice continuum is developed, with four stages: research to development to diffusion to practice. Evaluation serves as a possible method for getting researchers and practitioners to cooperate. Two major points are discussed. (1) The concept of evaluation is changing rapidly and becoming more pervasive; (2) the methodologies currently in use are hopelessly bad and urgently need replacement. Traditionally, evaluation has involved the comparison of some output with a set of absolute standards, and the comparison of two or more methods of doing the same thing in a relative sense. Measurements taken to carry out these classic forms of evaluation are usually of the pre- and post-test type. Generally, the rules of experimental design and field control are invoked and the task is to judge. Emergent evaluation, however, is seen as a process of collecting and interpreting data relevant to a series of decisions which must be made to aid in decision-making. Traditional evaluation has four major limitations: terminal availability of data, retrospective view, imposition of constraints, and limited generalizability.

The new kind of evaluation should probably have the following characteristics: (1) controls cannot be the typical laboratory controls but must be appropriate to the field. (2) Data collection must be carried on without disturbing the situation or the subjects. (3) Data must be collected continuously. (4) Treatments must be susceptible to change. (5) Attention must be given to any variables which appear to be of concern. (6) The assumptions of the evaluation must be formed to meet the reality of the situation and not vice versa.

Guba, Egon G. "The Failure of Educational Evaluation," *Educational Technology*, vol. 9, No. 5 (1969), 29–38.

This paper is a critique of much current evaluation practice, including the use of experimental design for evaluation. It urges more flexible designs that yield information more useful for program planning and modification.

Meaningful evaluation is difficult because of seven basic lacks. (1) There is a lack of an adequate definition of evaluation. The "measurement" orientation is too narrow; the deter-

mination of congruence between objectives and performance leads to the operationalization of goals in overly simple behavioral terms; professional judgments alone are too uncertain and ambiguous. (2) There is a lack of adequate evaluation theory. The classical experiment is usually inappropriate because it deals with "antiseptic" conditions rather than the septic real world, it yields information only after the program is completed, and it deals with one treatment at a time. (3) There is a lack of knowledge about decision processes and information requirements. Yet evaluation is expected to provide information for decisions. (4) There is a lack of criteria by which to judge results. The same data, interpreted according to different criteria, can lead to very different conclusions. (5) Differentiation of approach by levels is lacking. Clearly the focus of an evaluation of a classroom will be different from that of a state school system. Summaries of micro-level data do not always reflect the state of affairs, or meet the needs, of the macro-level. (6) There is a lack of mechanisms for organizing, processing, and reporting evaluative information. (7) There is a lack of trained personnel.

Efforts to cope with these deficiencies must be operationalized quickly.

Guba, Egon and John Horvat, "Evaluation During Development," *Bulletin of the School of Education, Indiana University,* vol. 46, No. 2 (1970), 21–45.

Current practices of evaluation in education are severely deficient. Educators concentrate too much on just measurement or on just one behavior, and their evaluations do not accurately reflect the actual program. A proposed approach defines evaluation as the process for obtaining and providing useful information for making educational decisions. There are four types of decisions—planning decisions concerned with ends or goals, structuring decisions concerned with means and implementation, implementing decisions concerned with utilizing procedures, and recycling decisions concerned with modifying or terminating the activity. For each type of decision a different type of evaluation is appropriate. Context evaluation, to define the environment and identify the problems, aids in making planning decisions. Input evaluation which analyzes procedures in terms of

costs and benefits helps to make structuring decisions. Process evaluation, to provide feedback on the success or failure of current procedures, assists implementing decisions. Product evaluation, to measure and interpret attainments and outcomes, is used to make recycling decisions. The authors give an example of an agency with a federal grant to alleviate the educational problems of migrant farm workers to show how each kind of evaluation leads to each kind of decision. They also show that evaluation and the development of the program should go on simultaneously, evaluation providing essential information at every stage.

Guba, Egon, G., and Daniel L. Stufflebeam, "Evaluation: The Process of Stimulating, Aiding and Abetting Insightful Activity." Address delivered at the Second National Symposium for Professors of Educational Research, November 21, 1968. Columbus, Ohio: Evaluation Center, College of Education, Ohio State University, 1968. (*Unpublished*)

There is at present a growing need and demand for evaluations, but educational evaluations have been poor in quality. Impediments to quality evaluation are: (1) inadequate previous definitions of evaluation, which have focused on measurement, goal achievement, or professional judgment; (2) poor classifications or conceptualizations of the variety of educational settings; (3) inadequate formulation of decision processes and informational requirements; (4) acceptance of experimental design as the ideal, whereas it is not appropriate to most educational evaluations (The classical experimental design conflicts with the principle that evaluation should continually feed back information for program improvement; it is useless as a device for making decisions in the planning and implementation stages; it is not suited to the septic conditions of the classroom, so that internal validity is obtained at the expense of external validity, i.e. generalizability.); (5) poor understanding of the kinds of evaluation resources needed; (6) no commonly accepted criteria for judging the quality of evaluations.

Educational evaluation is defined as the process of obtaining and providing useful information for making educational decisions. A

typology of educational decision settings is developed along two major dimensions: "small" versus "large" educational change and "high" versus "low" understanding to support change. These two dimensions lead to four major types of decision making settings, each of which requires an evaluation which is geared to the provision of the relevant type of knowledge.

Four different styles of decisions within each decision setting are also identified: planning decisions concern identification of objectives; structuring decisions refer to the design of specific procedures; implementing decisions are relevant to the utilization, refinement, or control of procedures; and re-cycling decisions concern judgment about the attainment of the program goals. Based on these definitions are four general evaluation designs, which form the CIPP evaluation model. For planning decisions, *context* evaluation is appropriate; for structuring decisions, an analysis of *input;* for implementing decisions, *process* evaluation; and for recycling decisions, *product* evaluation. Each of these evaluation types is described.

The main stages of all evaluations are: (1) focusing the evaluation, (2) collecting the information, (3) organizing the information, (4) analyzing the information, (5) reporting the information, (6) administering the evaluation. Criteria by which all evaluations may be judged are: (1) internal validity, (2) external validity, (3) reliability, (4) objectivity, (5) relevance, (6) significance, (7) scope, (8) credibility, (9) timeliness, (10) pervasiveness, (11) efficiency.

Gurin, Gerald, *Inner-City Negro Youth in a Job Training Project: A Study of Factors Related to Attrition and Job Success.* Ann Arbor, Michigan: Survey Research Center, 1968. (*Unpublished*)

This is a report of a research study of the JOBS—I project, an experimental and demonstration job training project for approximately 1,500 underemployed inner-city Negro youths in 1963–1964. The study staff interviewed a sample of almost 400 trainees at least twice—at the time of entrance into the program, during its course, and/or at termination. Interviews were also held with mothers of the trainees, supervisors on their first post-program jobs, and the JOBS project administrators, teachers, and counselors.

Since there was no control group, no attempt was made to estimate the benefits occurring from the program. Rather the study examined the characteristics and attitudes that differentiated the more successful from the less successful trainees.

The two major success criteria were retention in the program and job earnings in the post-program period. However, the author questions the use of retention-graduation as a success criterion. Graduates had more regular employment than dropouts, but their pay rates were no higher. Job earnings were related to better pre-program history, more education, and (for males) a higher proportion of employed males in the family. Earnings were also related to situational factors such as age and residence in own rather than parents' household, reflecting increased pressures to hold a job because of growing family responsibilities. The only attitude measure related to earnings was a "personal efficacy" scale designed to measure the sense of effectiveness and control; the relationship was positive and striking for male trainees.

Recommendations include more emphasis on job placement and job development, more stress on training in actual work settings rather than on preparatory training, focusing counseling efforts on problems as they arise on the job, making the training realistic as well as supportive.

Hagen, Elizabeth P. and Robert Thorndike, "Evaluation" in *Encyclopedia of Educational Research,* Third edition, pp. 482–486, New York: Macmillan, 1960.

The article is concerned with evaluation used for the guidance of students or used to assess some aspect of the curriculum. The process of evaluating involves three distinct aspects: (a) selecting the attributes important for judging the worth of the student or program to be evaluated and defining objectives, (b) developing and applying measures which will accurately describe the attributes of the student or program, especially self-evaluation techniques, and (c) synthesizing the evidence yielded by these procedures into a final judgment of worth. Research needs in education are identification of significant outcomes, improved devices for measuring student behavior, and

new ways to integrate results of measures into a comprehensive evaluation.

Hall, Richard H. "The Applied Sociologist and Organizational Sociology," in *Sociology in Action,* edited by Arthur B. Shostak, pp. 33–38. Homewood, Illinois: The Dorsey Press, 1966.

This case study of an evaluation of communication and coordination processes in an educational unit points out that research is often commissioned with no firm commitment to utilize the results. The research project was effectively and rapidly completed, administrators were actively involved in the formulation of the project, and effort was expended in meeting with the unit heads to explain in detail the findings of the study. Despite this, all recommendations, except those which concurred with the pre-judgments of the administrators, were ignored. The clear conclusion may be drawn that before time and effort are given to requests for research, the sincerity of the user organization should be established.

Hammond, K. R. and F. Kern, *Teaching Comprehensive Medical Care.* Cambridge: Harvard University Press, 1959.

In 1953 the University of Colorado School of Medicine initiated a general medical clinic program to teach fourth year students the techniques and philosophy of comprehensive medical care—the responsibility for a patient's total health and recognition of the importance of social and psychological factors combined with awareness of preventive techniques. The three year study (1954–56) was designed to (1) give the student the maximum possible responsibility for his patients, (2) increase continuity of contact with patients, (3) incorporate preventive techniques in clinical teaching, (4) stress importance of family-interpersonal relations, and (5) stress the importance of social and psychological problems in medicine.

Half of each senior class of 80 was assigned to the program for 24 weeks. Equal numbers from each academic third of the class were assigned to experimental and control groups, the GMC program and the usual clinical clerkship. Nine dependent variables were examined —knowledge, skills, and attitudes in medicine, sociology, and psychology. Pre- and post-tests on filmed doctor-patient interviews requiring application of knowledge and skill in all areas were administered. Students also took medical attitude tests on social aspects of medicine.

The program did not greatly affect the acquisition of either medical knowledge or awareness of psychological and social components of medicine, or increase skill in applying such knowledge. The experimental student was not more inclined to deal with psychological problems. His attitude toward comprehensive care remained the same, while that of the control students became increasingly negative.

Hansen, Morris H., William N. Hurwitz, and William G. Madow, *Sample Survey Methods and Theory,* vol. 1: *Method and Applications.* New York: John Wiley and Sons, 1953.

Topics covered in this basic text on sampling include: sampling principles; bias and non-sampling errors; sample designs for common sampling problems; simple and stratified random sampling; simple one or two-stage cluster sampling; stratified single or multi-stage cluster sampling; control of variation in size of cluster in estimating totals, averages, or ratios; multi-stage sampling with large primary sampling units; estimating variance; regression estimates, double sampling, sampling from time series and systematic sampling; and several case studies. There is an emphasis on practical application—such as cost factors and simple rules for approximating the optimal sample design—as well as on basic principles.

Hardin, Einar, and Michael E. Borus, "An Economic Evaluation of the Retraining Program in Michigan: Methodological Problems of Research," *Proceedings of the Social Statistics Section,* American Statistical Association, 1966, 133–137.

This article reviews problems of data collection in the analysis of the social economic costs and benefits of retraining programs, and makes recommendations for future studies.

A cost-benefit analysis that infers social product gain from the differences in earning between trainees and non-trainees tends to overstate the social economic benefits in periods of large general unemployment and understates the benefits in periods of labor shortages in certain occupations. If an analysis is to be made of the immediate effects of retraining programs, more weight should be given to results based on labor markets where unemployment is fairly low and evenly distributed, rather than where a general recession is evident. The interview method of collecting data on private earnings is inaccurate and expensive, and it usually gives incomplete results. Other methods of data collection in this area should be developed. Finally, the measurement of social economic costs should use a wider variety of indicators, such as the capital costs of instruction, operating costs, and the dependence of overall administrative costs upon the number and nature of retraining courses undertaken.

Harris, Chester W. (Ed.), *Problems in Measuring Change.* Madison: University of Wisconsin Press, 1963.

Measures of change using pre-test, post-test differences are less reliable than the scores from which they are derived. This volume is a collection of articles dealing with various procedures and models for analyzing change that attempt to cope with this unreliability. A fairly sophisticated understanding of statistical concepts is needed in reading most of the papers. The following articles are included:

"Some Persisting Dilemmas in the Measurement of Change," Carl Bereiter

"Elementary Models for Measuring Change," Frederic M. Lord

"The Reliability of Changes Measured by Mental Test Scores," John Gaito and David E. Wiley

"Multivariate Analysis of Variance of Repeated Measurements," R. Darrell Bock

"Multivariate Models for Evaluating Change," Paul Horst

"Implications of Factor Analysis of Three-Way Matrices for Measurement of Change," Ledyard R. Tucker

"Canonical Factor Models for the Description of Change," Chester W. Harris

"Image Analysis," Henry F. Kaiser

"The Structuring of Change by P-Technique and Incremental R-Technique," Raymond B. Cattell

"Statistical Models for the Study of Change in the Single Case," Wayne H. Holtzman

"From Description to Experimentation: Interpreting Trends as Quasi-Experiments," Donald T. Campbell.

Hastings, J. Thomas, "Curriculum Evaluation: The Why of the Outcomes." *Journal of Educational Measurement,* vol. 3, No. 3 (1966) pp. 27–32. Also reprinted in *Readings in Measurement and Evaluation,* edited by Norman Gronlund, pp. 53–60, New York: The Macmillan Company, 1968.

Evaluation serves two functions in curriculum innovation: the feedback of information to stimulate further innovation in curriculum, and the provision of information on which to base decisions about adoption of course-content packages. Different kinds of data are needed for the purposes of revision and adoption. Suitable standardized tests may provide enough information for adoption decisions, but statistical averages and summaries will give only vague hints about what needs to be revised. Examples of on-going research which would be appropriate for this latter function are given, e.g. research on study modes that benefit different types of students. Such projects straddle the line between pure and applied research: they investigate general issues in instructional research, but they do so in the context of specific curricula. They can therefore help to answer the "why" questions about curriculum outcomes.

Havelock, Ronald G. *Planning for Innovation through Dissemination and Utilization of Knowledge.* Ann Arbor: Institute for Social Research, University of Michigan, 1969.

This large volume is based on a comparative study of the literature on dissemination and utilization of scientific knowledge. (The extensive bibliography is bound separately.) It provides a framework for understanding the processes by which knowledge moves from the "resource system" to the user. Major sections analyze characteristics of individuals and organizations that inhabit or facilitate this transfer.

Three principal models of dissemination and utilization are examined: (1) research, development, and diffusion, (2) social interaction, and (3) problem solving. A fourth model, the linkage model, is developed to incorporate important features of all three. Linkage is the important interaction between user and resource systems

that culminates in mutual trust and the appreciation of each system's needs, patterns, and processes. Factors that help to explain dissemination and utilization are identified: linkage, structure, openness, capacity, reward, proximity, and synergy.

The report concludes with recommendations for needed research and development on the dissemination and utilization process and lays down guidelines for practitioners and government policy makers.

Hayes, Samuel P., Jr. *Evaluating Development Projects: A Manual for the Use of Field Workers,* Paris: UNESCO, 1966. (1959 Title: Measuring the Results of Development Projects).

This volume endeavors "to demonstrate how certain social science measurements can be adapted to help field workers assess initial conditions before a project is begun; to measure the extent to which various attempts at producing social change have been successful; to determine the over-all result of social development schemes; and to identify factors that are important in influencing the success of programs of social change."

The tasks of the development planner are to determine changes which need to be speeded up, to design projects by government or other organizations which give most promise of effecting these changes with the highest benefit-to-cost ratio, and to administer the projects efficiently.

The four logical steps in identifying and measuring the changes which ensue are examined in detail:

(1) Describing the project and specifying the goals.

(2) Deciding which data to use to indicate changes in the direction desired.

(3) Collecting before, during, and after data.

(4) Analyzing and interpreting the findings and reviewing them with the interested groups.

A "statistical concepts and elementary procedures" appendix is included.

Hemphill, John K., "The Relationships between Research and Evaluation Studies" in *Educational Evaluation: New Roles, New Means,*

68th Yearbook of the National Society for the Study of Education, edited by Ralph W. Tyler, Chapter IX. Chicago: NSSE, 1969.

Two major areas are discussed in this paper: (1) the symbiotic interplay between pure research and evaluation studies and (2) statistical decision theory as an appropriate means of appraising evaluation studies.

The first area is elaborated through the use of a case study of an evaluation of a nursery school for underprivileged children. Four goals of the school were to be evaluated: improvement in the child's self-image, development of his sensory and perceptual acuity, improvement of his language abilities, and development of his conceptual and problem solving abilities. In describing the process of evaluation, the author shows how the knowledge needs or discoveries of applied research feed back into pure research. In the case of the first goal, the evaluators discovered that there was no satisfactory way of assessing young children's self-image and the author suggests that research is needed to clarify the basic concept. Development of sensory and perceptual acuity was somewhat less difficult to operationalize, but questions were raised about whether the dimensions tested (color recognition and matching) were the most relevant to development, a question which they were unable to answer from the literature. In gathering data on language ability, problems of analysis arose because much of the data gathered was of a case or clinical nature, and the evaluators urge pure researchers to direct their attention to studies of these problems. Finally, in testing for concept formation and problem solving, no tests were found which related easily to the objectives of the school. A few "homemade tests" were used with great success (i.e. correspondence with intuitive expectation) and it is suggested that such efforts be elaborated in laboratory conditions.

The use of statistical decision theory to decide when it is worthwhile to do an evaluation is illustrated in a hypothetical case of a school principal who wishes to persuade a skeptical school board of the worth of a new program. With minimal knowledge about the cost of the program per child, and an intuitive sense about the probable success of the program and about the possible outcomes of the proposed evaluation, the principal is able to generate a set of

prior probabilities of the *costs* to the school for every possible outcome of every possible decision sequence: install the program with no evaluation, install the program with an evaluation that shows positive results, install the program with an evaluation that is negative, etc. He is thus able to recommend to the school board that an evaluation of the program be made since he can logically show that the eventual probabilities of high benefit-cost rates are measurably increased.

Herman, Melvin, "Problems of Evaluation," *The American Child,* vol. 47, No. 2 (1965), 5–10.

The author states that youth work programs have rarely been evaluated systematically. Based on a national study, he concludes that current evaluations suffer from deficiencies in: the definition of program objectives, use of adequate indicators of success, the collection of accurate and sufficient data, and the generation of valid conclusions. Evaluators often lack detailed understanding of the program and the milieu (in this case, specifically the labor market) within which it operates. Herman illustrates his discussion with accounts of the Mobilization for Youth work program which he formerly directed.

Herman, Melvin and Michael Munk, *Decision Making in Poverty Programs: Case Studies From Youth Work Agencies,* New York: Columbia University Press, 1968, pp. 139–181.

The last section of this book consists of five case studies illustrating problems of research and evaluation which commonly arise in poverty programs and similar situations. The emphasis is on policy issues and interpersonal relations, rather than on methodological or theoretical questions.

The first case describes a situation in which a research department found itself responding to essentially political rather than professional issues. This illustrates the potential vulnerability of the research department to organizational needs. The second case deals with the conflicting needs for hard data for purposes of program feedback, as compared with longer-term evaluation. In the third case, the authors deal with the problems that outside evaluators face in gaining access to and disseminating informa-

tion, problems which arise because of differing expectations as to the nature of the evaluation. The fourth case examines an attempt to convert an existing service activity into a controlled experiment. The difficulties encountered in this case highlight the potential value differences between the practitioner and the academician. In the last case the issue was designing an evaluation prior to the implementation of the program. Straight methodological problems are compounded by "in house" fears of "outside" scrutiny. Questions about crucial decision-making stages in each case study are presented.

Herzog, Elizabeth, *Some Guide Lines for Evaluative Research,* Washington, D.C.: U.S. Department of Health, Education and Welfare, 1959.

This clear, well written booklet is concerned with the measurement of psycho-social change in individuals, and is organized around nine important issues in evaluation research: (1) The purpose of the evaluation. What will be achieved by it? (2) The kind of change desired. Answering this question will involve determining the original state, what change is desired, what criteria will be used to indicate change, and identifying the group in which change is expected and its outstanding relevant characteristics. (3) The means by which change is to be brought about. The evaluator here must differentiate between theory and practice, and must determine who the change agents are and how they have affected the process. (4) The trustworthiness of the categories and measures employed. Assessment of reliability and validity are essential if the evaluation is to carry weight. (5) The points at which change is to be measured. Baseline measurements, location of a sample, choice of interviewers, and intervals between base and final measures are discussed under this section. (6) The representativeness of the individuals studied. Selection and definition of the sample will affect the generalizability of the results. (7) Evidence that the changes observed are due to the means employed. Problems of establishing adequate controls in social research are covered, and suggestions are given. (8) The meaning of the changes found. (9) The unexpected consequences of the program or treatment.

The points raised by the author are illus-

trated from an extensive list of references to actual studies in psychotherapy and social work.

Hessel…ing, P. "Principles of Evaluation," *Social Compass*, vol. 11, No. 1 (1964), 5–22.

Three dimensions of evaluation research are identified and described: assessment for whom, assessment by whom, and the time and stage at which the evaluation is conducted. A typology of evaluation is presented which integrates the dimensions "for whom" and "by whom", and 25 types of evaluation are defined. Four stages of evaluation are distinguished: (1) the determination of general objectives, which should occur early in the program's development, (2) the assessment of specific objectives and needs, which should also be undertaken before massive implementation, (3) observations and recordings of program activities, in order to determine whether the program is doing what it is supposed to, and what might be improved, and (4) assessment of outcomes and effectiveness. These four stages are all important if evaluation is to become an integral part of action.

Hill, Marjorie J., and Howard T. Blane, "Evaluation of Psychotherapy With Alcoholics," *Quarterly Journal of Studies on Alcohol*, vol. 28, No. 1 (1967), 76–104.

Forty-nine studies are reviewed and compared on the basis of five criteria: (1) use of controls, (2) subject selection procedures, (3) selection and definition of criterion variables, (4) measurement instruments and their reliability, and (5) measurement before and after treatment. It is concluded that almost all the studies are relatively worthless, because of their methodological deficiencies.

Hochstim, Joseph R. "A Critical Comparison of Three Strategies of Collecting Data from Households," *Journal of the American Statistical Association*, vol. 62, No. 319 (1967), 976–989.

This article compares three strategies of data collection: personal interviews, telephone interviews, and mail questionnaires. The three strategies were tested on area probability samples of households in Alameda County, California. Two separate studies were made, with identical questionnaires used in all strategies within each study. The responses from the three

strategies were found to be highly comparable. Rate of return and rate of completeness of answers were high for all three, substantive findings were virtually interchangeable, and there was little difference in validity. The three strategies did vary considerably by cost, however. Personal interviews are the most costly.

Holliday, L. P., *Appraising Selected Manpower Training Programs in the Los Angeles Area*. Santa Monica, California: Rand Corporation, May 1969.

This report summarizes the principal theoretical and empirical findings of a project conducted for OEO to develop methods of evaluating manpower training programs. As part of the project, an exploratory study was conducted of former enrollees in a training project in Los Angeles and employers of program graduates. This was supplemented with observations of counseling and classroom interactions and project cost analysis based on records and staff interviews.

Major recommendations include: longitudinal study; development of proximate measures of program outcome (in addition to job placement rates) such as changes in attitudes, motivations, economic welfare, etc.; use of low-cost sources of follow-up data, such as employer surveys; computer-based information systems; development of a set of standards for cost-benefit analysis; better understanding of the decisions for which the evaluation provides data; more use of multivariate analysis, such as regression analysis; more work on the control group problem; analysis that distinguishes program effects from population characteristics; more formalized procedures for program development.

Holmstrom, Engin I. and Laure M. Sharp. *Study of NDEA Title IV Fellowship Program, Phase II*. Washington, D.C.: Bureau of Social Science Research, publication #397, July 1970. (*Unpublished*)

The study was designed to evaluate the effectiveness of the NDEA Fellowship Program (which provides financial aid to graduate students for up to three years) in facilitating completion of the doctorate and increasing the number of college teachers. This phase of the study used mail questionnaires to recipients of

---

fellowships in 1960–62 and to a comparison group of graduate students. Questionnaires were sent to 2,983 NDEA Fellows and 1,141 non-recipients in 1969.

Problems encountered in the study were: (1) all Fellows were studied, although 18% had dropped out prior to completion of the three-year period, (2) the awarding process singled out the more promising students and made a rigorous control group impossible, (3) the comparison group had to be constructed retrospectively through nominations of students by deans; it was not truly comparable, (4) there were problems in obtaining addresses and securing adequate returns. The evaluation thus yields suggestive insights rather than firm conclusions.

A majority of both NDEA Fellows and comparison group students had earned the doctorate by the time of the study. Comparison group doctorates took longer to obtain the degree than NDEA doctorates. Similar proportions (about two-thirds) of both groups were holding academic positions. Thus, since NDEA fellowships "allow a large number of graduate students to obtain the doctorate in a relatively short period of time, and since a majority of the doctorates become teachers, one might say that the program was successful in its objective of increasing the number of qualified college teachers."

Hough, Robbin R. "Casualty Rates and the War on Poverty," *American Economic Review*, vol. 58, No. 2 (1968), 528–532.

This paper focuses on the inadequacy of data collection processes in large scale action programs. An alternative, more systematic data collection program is proposed, which involves gathering three different types of data to fulfill three different functions: (1) to demonstrate that a problem exists on a scale large enough to warrant attention; (2) to demonstrate that the funds were legally spent; or (3) to estimate the impact which a program had while it was operating. At present, action programs yield little public data by which to evaluate the interaction between institutions generated by the program and to gauge the progress of the program toward its goals. In addition, lack of factual information inhibits the important potential dialogue between academics and administrators, which might prove

of service for short run developments. A review by academics of day-to-day decision-making at all levels of the administrative process might result in the integration of systematic models into the routine administrative data collection process. If data collection efforts were rationalized along model-building lines, the data banks produced would be useful in generating fruitful interaction between scientist and practitioner.

Houston, Tom R., Jr. and Julian C. Stanley, "The Behavioral Sciences Impact Effectiveness Model." Paper presented at Evaluation of Social Action Programs Conference, American Academy of Arts and Sciences, May 2–3, 1969. (*Unpublished*)

This paper states the case for the experimental model, with randomized experimental and control groups, as the optimal design for evaluation of social programs. It discusses the utility of factorial design to isolate the effects of specific components of programs for specific participant groups.

Hovland, Carl I. "Reconciling Conflicting Results Derived from Experimental and Survey Studies of Attitude Change," *American Psychologist*, vol. 14, No. 1 (1959), 8–17.

The paper discusses the two types of research which study the modification of attitudes through communication—the controlled experiment and the survey method which uses correlations between reports of exposure and measurements of attitude. The conclusions derived from the two methods are often divergent. Correlational studies often show little effect of communication, and experiments tend to show considerable effect.

The critical variation between the designs is the difference in definitions of exposure. In the naturalistic survey the audience must expose themselves to communication, while in an experiment exposure is enforced. In an experiment the effect is observed directly after exposure; in a survey more remote effects may be measured. The types of communicators and the motive-incentive conditions are different in the two designs, increasing the likelihood of change in the experimental study. Populations used and types of issue also are different between the two types of studies.

A second main source of disparate results is the varying distance between the position taken by the communicator and that held by the subject. The difference here is probably due to the differential involvement of the subject with the issues. Surveys often deal with more important, basic issues. A third area is the contrast between the naturalistic survey's emphasis on primacy in order of presentation with the not-very-significant effects of primacy found in the laboratory. The key variable here is the fact that self-exposed groups tend to examine only one side of the issue while an experimental subject is given both sides.

Thus no contradiction has been established; the seeming divergence can be accounted for by differences in situation, communicator, audience, issue, etc. A genuine understanding of the effects of communication on attitudes requires both experimental and survey methodologies, as each offers an important emphasis.

Hutchison, George B. "Evaluation of Preventive Services," *Journal of Chronic Diseases*, vol. 11, No. 5 (1960), 497–508.

This paper discusses the methods and requirements for evaluating the effect of early discovery of disease in preventive medicine, as opposed to normal symptomatic diagnosis. The evaluation of early diagnosis is necessary because of conflicting views regarding the value of early detection. The general evaluation model, applicable to all preventive medicine programs, asks: (1) Does the program meet its objectives? (2) To what degree does it meet the objectives? or (3) How efficiently does it meet its objectives?

The ultimate objective of a preventive program is to alter the natural history of a disease in a favorable direction. In order to measure the success of a program, it is necessary to develop an analytic description of the natural history of the disease. From this analysis it is possible to determine whether early diagnosis, and thus early application of therapy, is beneficial. In order to predict benefit from early discovery, a disease must have the following characteristics: (1) There must be a known effective therapy. (2) There must be a diagnostic device capable of detecting the disease before the usual time of diagnosis. (3) There must be one or more critical points such that therapy ap-

plied before the critical point is more effective than therapy undertaken after the point. (4) A critical point must occur after the time when diagnosis first becomes possible and before the time when diagnosis is usually made. An evaluation of an early detection program must take into account outside factors such as the community in which the program is operating, physical conditions like air pollution, level of medical awareness, and the sophistication of both the public and the medical profession. Considering all factors, proof must be found of the maxim, "the earlier the treatment the more effective the prevention," in evaluating preventive medicine programs.

Hyman, Herbert H. and Charles R. Wright, "Evaluating Social Action Programs" in *The Uses of Sociology*, edited by Paul F. Lazarsfeld, William H. Sewell, and Harold L. Wilensky, pp. 741–782. New York: Basic Books, 1967.

This chapter is particularly useful in its attention to the definition of the program that is to be evaluated. The term "program" is deceptive, since it encourages viewing the subject of study as both an actuality and an entity. In fact, one of the first tasks of the investigator is to determine to what extent there is a discrepancy between the plan and the program. In evaluating a program which is extended in time and space, one may in fact be dealing with a variety of programs. In studying one cycle of a program, or even a limited program in a continuing organization, assessment must be made with at least some knowledge of the context of previous cycles or activities which have occurred. When cycles overlap, (for example, in the situation of a school) care must be taken to separate interaction effects from program effects.

In conceptualizing a program, the site and the staff must be considered as independent variables along with the actual treatment and its temporal context. Since evaluation research is concerned primarily with empirical testing, the process of conceptualization should be limited to significant variables which can be operationalized.

A major task is the identification of specific, critical goals which can serve as the basis for determining the program's relative success,

since outcome must be related to original intent. The major dimensions for conceptualizing the effects of a program are: (1) the locus of effects—whether the program goals are geared to individuals, communities, total societies, or a combination of these, (2) the temporal aspects of effects—expectations of developmental sequences for attitudinal or behavioral changes, persistence, whether social chains will carry the effects outside of the immediate target population, etc., and (3) unanticipated consequences.

The authors recommend that the design of evaluation research be comparative whenever possible. Comparisons may be made between factors within a program, between programs in the same setting, and between programs in different settings. Designs which are replicative or longitudinal in aspect also provide a better basis for determining long range effects of social programs. Hyman, Wright, and Hopkins' study of the Encampment for Citizenship is discussed in detail as an example of research involving continuity, replication, and longitudinal evaluation.

Hyman, Herbert H., Charles R. Wright, and Terence K. Hopkins, *Applications of Methods of Evaluation: Four Studies of the Encampment for Citizenship*. Los Angeles: University of California Press, 1962.

This is an evaluation of the effectiveness of the Encampment for Citizenship, a six week summer camp for young adults, whose goal was to prepare young American citizens for responsible citizenship and citizen leadership. The design of the study which was based primarily on the collection of survey data from participants, has several interesting features: (1) A major problem arose in attempting to conceptualize and operationalize the desired goals of the program. The behavior which the program hoped to affect was future, and could not be measured. Scales and inventories on attitudes which were theoretically connected to future behavior were therefore developed in seven basic areas: orientation toward civic activity, cognition of social problems, salient social attitudes and opinions, perceived relationships with the rest of society, certain skills and capacities, and present conduct. (2) The evaluation attempted to incorporate several experimental controls which would aid in a more

rigorous interpretation of the data. First, instead of taking only one baseline measurement before the beginning of the camp, two were taken: one six weeks before the camp, and one just before it began. Thus, information about the extent of normal variation on the attitude scales was obtained, and could be discounted in analyzing programs. Second, attempts were made to find similar groups of young adults who were not in the program, but who received alternative treatment. Participants in a volunteer work camp were also given before-and-after tests, and this allowed the authors to control for maturation and make some crude judgments about the relative effectiveness of the program. In later replications, those who were unable to attend the camp for some reason formed the control group. (3) The design included both a simulated and a real longitudinal study. Earlier graduates of the program were tested to determine long-range effects, and the original group studied was followed up four years later. Replications were also used to increase the cogency of the findings, which showed that the Encampment had a considerable effect on its participants and that the effect persisted over time.

James, George, "Planning and Evaluation of Health Programs," in *Administration of Community Health Services*, pp. 114–134. Chicago: City Managers Association, 1961.

This chapter is a discussion of the rationale and methods for planning and evaluating local health department programs. The planning stages serve to locate omissions and inefficiencies in the program, and to identify community needs, resources, and attitudes in order to define practical objectives. Once the plan has become a specific program, evaluation is necessary to measure the degree to which it is achieving its goals. Since many new health programs are expensive and of unproven value, evaluation is essential in aiding public health administrators in rearranging priorities among present and future programs. Evaluation can study (1) effort, (2) performance, (3) adequacy of performance, and (4) efficiency. The latter two types are most complex, and should not be attempted until after an assessment of the first two, although they provide the most useful information for planni....

Evaluation procedures should be built into the program, in order to provide feedback information relevant to program redesign, and to facilitate data collection in a more systematic way than is possible in a retrospective study. Evaluation is also needed in order to stimulate dissatisfaction with traditional programs, thus encouraging a critical analysis of their objectives, assumptions, and performance. Two methods for building evaluation into the program are suggested: (1) arranging for periodic reviews of the program by outsiders, (2) using interdisciplinary program teams, thus increasing the chances that all members of the team will not have the same perspective on the program's effectiveness.

Critical evaluation has often been given low priority due to community or interdepartmental preconceptions and traditions. The public health administrator must look for special opportunities (for example a period of tight budgets, or a sudden overwhelming health need) to put into effect the recommendations which grow out of evaluation studies.

James, George, "Research by Local Health Departments: Problems, Methods, Results," *American Journal of Public Health*, vol. 48, No. 3 (1958), 353–361.

In this discussion of the types and importance of research done by public health departments, evaluation research is seen as particularly important to the local health officer. Evaluation of traditional programs in school health, immunization, and general sanitation, can be used to cut wastes of money and energy. For example, only evaluation can be used to find whether it is really necessary to maintain elaborate milk control programs to insure high quality of milk. Evaluation also helps to identify the best ways to allocate resources and funds. For example, a study of school nurses revealed that these nurses spent most of their time in low priority or nonprofessional activity. By adding nurse assistants, the public health nurses were freed for more professional duties. Operations research, a branch of evaluation, uses an interdisciplinary approach to the reduction of discrepancies between research findings and actual practice. It shows that evaluation results should not be looked on as mandates, but as available facts to be used when needed. Op-

erations research also tempers evaluation by assigning a value to human factors such as good will. Other types of research are also discussed. The author concludes that research is essential in maintaining effective, timely, and streamlined public health programs.

Justman, Joseph, "Problems of Researchers in Large School Systems," *Educational Forum*, vol. 32, No. 4 (1968), 429–437.

The problems of small staffs and myriad demands often make it difficult for research bureaus in school systems to do adequate research. Getting research done is not an easy task, and many problems must be faced, such as: (1) problems of priorities—what is most valuable to study, (2) organizing an integrated research program in spite of different needs of each part of the school system, (3) shortage of required personnel, (4) efficient, effective allocation of time and energy, (5) accessibility to the "powers-that-be," (6) the tendency of a school-based researcher to become parochial and narrow in his interests, (7) restricted publication of research findings and possible administration censorship, (8) uncooperative school personnel, (9) rapidly changing populations in the schools, (10) difficulties in getting equipment, especially computer-time for data processing.

Kahn, Robert L. and Charles R. Cannel, *The Dynamics of Interviewing*. New York: John Wiley and Sons, 1957.

This excellent text views the interview as a communication process, and covers such topics as the psychological basis of the interview, techniques for motivating the respondent, the formulation of objectives and questions, the design of questionnaires, the interview as a method of measurement, the use of probes, and learning to interview. It includes examples of interviews from medicine, personnel work and social work.

Kandel, Denise B. and Richard H. Williams, *Psychiatric Rehabilitation: Some Problems of Research*. New York: Atherton Press, 1964.

The book is based largely on a 1959 conference attended by representatives of 49 research and demonstration projects involved in

the rehabilitation of mental patients. Two-thirds of the projects were engaged explicitly in evaluation. The problems of conducting research in mental hospitals and allied institutions are analyzed in a sociological framework. Innovative demonstrations tend to disrupt the equilibrium of the existing social system which depends on rules defining the obligations of interacting persons toward each other and adequate motivation for people in the system to fulfill the obligations. The projects often interfered with one or both requirements by disrupting (1) the personality system of the members, (2) the structure of the social situation, or (3) the general cultural climate in which they were operating.

At the personality level, projects were perceived as threatening by practitioners, led to negative self-images, and did not provide adequate rewards. At the level of the social system, the projects (a) created new roles for clinical and research personnel, (b) did not clearly define the roles (even when not new) of clinicians and researchers, (c) placed people in conflicting roles, (d) led to conflicts between persons in different roles, (e) required cooperation from people with conflicting frames of reference, expectations, and perceptions, (f) often were deficient in communicating information to persons concerned, and (g) experienced problems with lines of authority and delegation of authority. At the level of the cultural system, conflicts in values and traditions developed.

Methodological problems arose, often as a result of the operational problems listed above. Particularly critical were the formulation of the research question in terms that were not too broad or vague, securing needed data, and establishing and maintaining control groups.

The authors offer suggestions for more successful action research: minimize disruptions, provide adequate structure (values, goals, and roles), improve communication, involve everyone who will be affected by the project early in the operation, be flexible, start with a pilot phase before going into more elaborate design and data collection, give more attention to basic assumptions and theoretical concepts, anticipate problems.

Katz, Irwin, "Review of Evidence Relating to Effects of Desegregation on the Intellectual

Performance of Negroes," *American Psychologist*, vol. 19, No. 6 (1964), 381–399.

This paper shows how an examination of theory can lead to new hypotheses about the effects of desegration on Negro children in the classroom. Psychological theory does not provide any unequivocal predictions. On the one hand, the child is likely to perceive whites as a social threat, and indifference or hostility from his peers might produce anxiety. On the other hand, studies have shown that if an individual is accepted in the group, he will adhere to the norms of the group, in this case higher academic performance. However, if these norms are substantially higher than those he is used to, he may become discouraged, and the probability of his failure will be high. Fear of disapproval should increase as it becomes more probable.

Using the results of past studies, the author attempts to isolate process variables which might affect adjustment. There are few reports on the performance of Negroes in desegregated schools. What is reported presents a favorable picture of Negro children's adjustment. Some evidence exists on desegregation conditions that may be detrimental, however: (1) conditions of social rejection and isolation, documented in reports and studies, may cause physical and psychological distress symptoms, (2) fear of competition with whites may exist, (3) inadequacy of previous training may interfere with adjustment, and (4) unrealistic inferiority feelings have been found in several situations. Experiments on stress and performance show that an organism's vulnerability to stress depends on the nature of its social environment. Isolation is one of the social environments which seem to render organisms most vulnerable to stress.

The author finally reports on his own experimental evaluations of Negro performance in biracial teams, where the participants have been matched for intellectual ability. One experiment showed that Negroes are more passive and compliant, rate their own performance as inferior even when it is not, and get less enjoyment from the team experience. A second experiment showed that when the situation is seen as non-threatening, performance is high, while the introduction of threat lowers it. In all cases,

however, Negroes performed better in racially homogeneous groups.

On the basis of experimental support for the theoretical framework, the author makes several recommendations for educational policy, including abolishing the track system, raising the level of predominantly Negro schools, and introducing desegregation at younger ages.

Kellner, Robert, "The Evidence in Favour of Psychotherapy," *British Journal of Medical Psychology*, vol. 40, No. 4 (1967), 341-358.

Past surveys of the results of psychotherapy have "shown" that it has little or no effect. In this article, nine studies which show conflicting results are summarized for the purpose of discussing methodological obstacles which may prevent the detection of measurable treatment effects. Three treatment groups are covered: children, adults, and juvenile offenders. In each area a study showing no effects is discussed, and then two with favorable results are presented. In all three areas, the author shows that the results of the unfavorable study are likely to have been affected by inadequate methodology.

Controlled studies tend to show that patients differ in their amenability to psychotherapy, but this is often not taken into consideration in the evaluation of psychotherapeutic results. Reviews of the effects of psychotherapy also ignore the evidence which suggests that different types of therapy are appropriate in different situations. Psychotherapy increases the variability of a treated group, and the comparison of mean scores may hide changes which have occurred: some patients may be harmed by inappropriate methods, and their conditions may become worse, while those in the control group remain the same. An important source of failure to detect changes may also be traced to the heterogeneity of the samples. This is particularly crucial where the effect of treatment is small in relation to the other factors. Those studies which show positive results have tended to control at least to some extent for variability among the sample group.

Kelman, Howard R. "An Experiment in the Rehabilitation of Nursing Home Patients", *Public Health Reports*, No. 77 (April 1962), 356-366.

This study, conducted by two departments of New York Medical College, examined rehabilitation services to determine whether the benefit derived from the services can justify their wider application to nursing home populations. Patients assigned to one of the two treatment groups were treated in the nursing home by a mobile rehabilitation team. The team devised and carried out a therapeutic program for each patient. Primary nursing responsibility was retained by the nursing home staffs. Consultation services were obtained from those who had normal medical responsibility for the patients. Patients assigned to the other treatment group were transferred from the nursing home to one of 5 rehabilitation hospital centers.

Over 2000 patients in 15 nursing homes in New York City, all welfare recipients, were reviewed. All patients with physical impairments that limited functioning in one or more self-care areas were included in the study. Two treatment groups and a control group, each composed of approximately 100 patients, were drawn from 11 nursing homes and a second control group from another 4 homes. The control patients received the usual care and services.

All patients were tested for their initial levels of self-care and the tests were repeated after a year's treatment. Five indices were employed in measuring self-care status: self-locomotion, ability to get from one position to another, ability to feed oneself, ability to toilet oneself, ability to dress oneself. Comparisons showed that neither rehabilitation in a hospital nor in a nursing home significantly altered functional self-care status. The rehabilitation treatment programs failed to influence favorably hospitalization and mortality.

Kelman, Howard R. and Jack Elinson, "Strategy and Tactics of Evaluating a Large Scale Medical Care Program," *Proceedings of the Social Statistics Section*, American Statistical Association, (1968), pp. 169-191.

The paper describes the attempt to develop a methodology for evaluating the impact on an urban ghetto community of the affiliation of the community general hospital with a University Medical Center. There are two major questions. First, how are the needs of the community met by the providers of health care, including the hospital, in the community? Second, how ap-

propriate and adequate is the care given by the hospital and what are its consequences for the recipients of care?

The first question can be answered by a survey of representative households in the community over time. The second requires a study of the hospital. The approach finally developed by the study leans heavily on a broadened concept of the medical audit and one one-year follow-up or outcome studies of selected groups of patients. The primary focus is the patient, his status during the course of hospitalization and subsequent to it.

Information gathered should help to answer the questions: (1) What is the character of the population now receiving hospital care? (2) What care needs are not being met? (3) To what extent would upgrading quality or quantity of care achieve different results?

This program is very complex and it is difficult to define the relevant evaluative questions. The system itself is not geared to the demands of evaluation research. The application of an experimental design is inappropriate or impossible. The Research models now available do not fit the needs for evaluation of this kind. As a result, an evaluation such as this has high risks in terms of immediate pay-off and eventual yield.

Kendall, Patricia, "Evaluating an Experimental Program in Medical Education," in *Innovations in Education* edited by Matthew B. Miles, pp. 343-360. New York: Teachers College, Bureau of Publications, 1964.

This is an evaluation of a medical education program whose objectives were to increase the amount of attention given by the students to social and psychological backgrounds among the patients, to increase student interest in the welfare of their patients without reducing the professional quality of the relationship, and to increase their sensitivity as to the kinds of practices that constitute quality medical care. All of these objectives were operationalized through lengthy discussions with the doctors who designed the program.

The faculty of the medical school refused to sanction a traditional experimental design which would have given one half of the senior class the treatment while the other half served as a control group. A modification of the ex-

perimental design was adopted, whereby one half of the fourth year class had the treatment the first semester, while the other half had it the second. It was also impossible to randomize, and although there were no indications of bias in the division of the groups, it was decided to use a panel design, where each student would serve as his own control. Thus, all students, whether they had the treatment course during the first or second semester, were measured at three points: at the end of the third year, the middle of the fourth year, and the end of the fourth year. This allowed the evaluators to assess not only the effects of the program in comparison with a control group, but to distinguish between long and short term effects. Other interesting aspects of the study included (1) the comparison of all the fourth year students with the first, second, and third year students so that natural trends in opinion change could be discerned, and (2) the attempt to include a comparison of the various elements of the experimental program, in order to specify the most effective factors in producing attitude change. Results indicated that the program had differential effects on students, depending on their original attitudes toward the role of the doctor. The author notes that there is a great need for replicative studies in the evaluation area.

Klineberg, Otto, "The Problem of Evaluation," *International Social Science Bulletin*, vol. 7, No. 3 (1955), 346-352.

The goals of Unesco programs have not changed but increasing scepticism must be directed at the methods used to reach them. It is crucial to develop methods of evaluation which are objective, systematic, and comprehensive. Administrators must be willing to accept the cost of evaluation and must be able to define strictly the goals and functions of their programs. A special problem arises in connection with international programs. Evaluators must be aware that a technique which works well in one culture may not work well in another. Also a knowledge of the culture is necessary in order to be alert to "unanticipated consequences" of action programs. Several critical surveys of evaluation methods and techniques by Unesco are briefly described.

Kogan, Leonard S. and Ann W. Shyne, "Tender-Minded and Tough-Minded Approaches in Evaluative Research," *Welfare in Review*, vol. 4, No. 2 (1966), 12–17.

This article deals with the evaluation of intervention programs designed to support the psycho-social functioning and welfare of individuals and families. Two distinct therapeutic approaches may be found: the "tough-minded" approach aims at behavioral modification, while the "tender-minded" approach stresses self-understanding, resolution of intra-psychic conflict, etc. Although these two approaches will produce different focuses for program evaluation (inasmuch as the goals are somewhat different), the evaluator should not limit his concerns to gross behavioral outcomes. Intrapsychic and intra-family interaction variables are also important.

Krause, Elliott A. "After the Rehabilitation Center," *Social Problems*, vol. 14, No. 2 (1966), 197–206.

This paper shows why simple follow-up data for the evaluation of a rehabilitation center do not constitute a valid evaluation. The article uses one small follow-up study conducted in a typical vocational rehabilitation program to show the sources of error and bias in the interpretation of simple follow-up data.

The data on vocational outcome after six months out of the center showed that 48.9% of the clients were working at the minimum wage or more, or in other words were "rehabilitated." The other 51.1% were "unrehabilitated" by standard criteria. However, many factors other than the program could have affected the "results." Diagnostic evaluation of a client's potential, only the first step of rehabilitation, was the major service of the center. If the center found that the client would not do well at any job, his case was closed. Counselors were under pressure from funding sources to place as many people as possible, and the severely handicapped were likely therefore to be refused service. Thus, the training group was not a representative sample of the clientele. Different agencies had different standards for the level of client performance at the center. Relations between the referring agencies and the center were often strained, preventing understanding of client progress, which led to un-

favorable outcomes. Client-family relationships were not under the control of the center but they had strong influence on the clients' eventual job placement or failure to find a job. At the time of the study, 1961–64, long-term unemployment among the lower class was increasing, confronting clients with real job shortages. The resistance of employers to hiring the handicapped was also a barrier to placing clients. These important intervening factors make it impossible to get accurate evaluation of the center's effectiveness by simply following the clients who pass through it. There is a need for more sophisticated research evaluations of such programs. Any substitute is a naive excuse for "evaluation" and cannot be credited with any validity.

Landers, Jacob, *Higher Horizons: Progress Report*, New York: Board of Education of the City of New York, January 1963.

The evaluation of the Higher Horizons program (a major educational program to help disadvantaged children by raising the levels of aspiration through special services, such as individual counseling, curriculum enrichment, and remedial courses), sought to answer three questions. (1) Does this program develop pupil potential more effectively than the conventional program? That is, is scholastic performance improved? Are there fewer disciplinary problems? Are attendance rates better? Is there better personality adjustment? Are higher but realistic vocational goals being sought? (2) To what extent, if any, does the use of a variety of methods and techniques, including teacher and counselor observations and ratings, identify more potentially able students among the deprived pupil groups than usual standardized tests of aptitude and achievement? (3) To what extent does parent and community participation in the experimental program result in raising the level of aspiration of educational and vocational plans of their children?

The evaluation was conducted in two parts, a normative survey and an intensive experimental-control study. For the normative survey a modified longitudinal approach was used to follow the growth and development status of 1,000 third grade and 1,000 seventh grade students for two or more years. The experimental-control study examined samples of third and

seventh grade students in the participating experimental and control schools, a total of 1,000 students. Academic achievement, personal adjustment, self-concept, unmet needs, conduct, and attendance rates were compared using both quantitative and qualitative approaches. School population, age, regular and special services, ethnic composition, pupil transiency, teacher turnover, non-English speaking population, attendance rates, school size, and class size were all examined. Data were obtained through questionnaires, sociometric techniques, inventories, checklists, and rating scales from teachers, students, counselors, and parents. (For results, see J. Wayne Wrightstone et al.)

LaSorte, Michael, "The Caseworker as Research Interviewer," *American Sociologist*, vol. 3, No. 3 (1968), 222–225.

The author argues that the use of social caseworkers as research interviewers in social action projects is doomed to failure because of the role conflicts inherent in this dual status. Using material from a project on which he worked, he shows how the lack of acceptance of sociological research norms undermined the research project. The values of the social work profession include immediate treatment and independent control over clients. Research demands were seen as challenges to both of these. Furthermore, social work is a profession with a psychological orientation and sociological formulations were less acceptable. Finally, and perhaps justifiably, the social workers resented the fact that they were required to accommodate their views to the researchers', while the researchers were unwilling to modify their research procedure at all. Both research and action staff therefore lapsed into antagonistic patterns, which discouraged convergence or cooperation.

Lemkau, Paul V. and Benjamin Pasamanick, "Problems in Evaluation of Mental Health Programs," *American Journal of Orthopsychiatry*, vol. 27, No. 1 (1957), 55–58.

Drawing on community research in the mental health field, the authors conclude that most programs currently being funded in this field are far too comprehensive for evaluation to show concrete results. More emphasis must be placed on questions that can be answered. Use

of genuine control groups and behavioral rather than opinion indicators will help evaluators to keep sight of the critical questions of the experiment. The vague nature of many mental health programs does not allow for measurable goals, much less measurable results.

Lempert, Richard, "Strategies of Research Design in the Legal Impact Study," *Law and Society Review*, vol. 1, No. 1 (1966), 111–132.

The paper discusses the application of experimental and quasi-experimental designs to the study of the effects of laws on behavior (legal impact studies). Legal impact studies must deal with a number of plausible rival hypotheses which could explain the change which the researcher wants to ascribe to the law. There is also the problem of distinguishing the law as it appears on the books from the law as it operates in fact.

Strengths and weaknesses of the experimental designs described by Campbell and Stanley are discussed. Design 1 is a simple descriptive design, with observations taken at one point in time. It does not allow any generalization. Design 2 calls for before and after observations of behavior which the law purports to regulate. It is better than 1, but it is weak in controlling for rival hypotheses and thus can be very misleading. Design 3 compares societies which have a particular law and societies which do not. This is useful for methods of thought rather than systematic research. Design 7 takes a series of behavioral observations at points before and after the passage of the law. This is weak because no control is used and several rival hypotheses, e.g. independent historical variables, are not controlled for. Design 10 calls for pre-test and post-test measurement from both experimental and control populations. It can lead to false or uninterpretable conclusions because of trend patterns, for example. Design 14 is the best for impact research. It resembles design 10, but many observations are made periodically before and after passage of the law in both experimental and control groups. This design rates very high on internal validity criteria and it best disposes of alternative explanations. However, if design 14 cannot be used, it is almost always better to use an inferior design than to do no study at all.

Lerman, Paul. "Evaluative Studies of Institutions for Delinquents: Implications for Research and Social Policy," *Social Work*, vol. 12, No. 4 (1968), pp. 55–64.

The author makes the innovative suggestion that *failure rate* may be a more appropriate criterion for evaluating many types of programs than *success rate*. Particularly in the case of treatment institutions, where success is difficult to define, demonstration of a reduction in failure rate would provide the necessary information.

The case of evaluating residential institutions for delinquent boys is examined. A major problem in determining the relative success of various types of treatments in this case is the high percentage of youths who are discharged from private institutions without "completing the treatment." The organizations prefer not to include them in an analysis of success-of-treatment rates, so that success rates appear higher. But those who fail internally and leave before the end of the treatment are "failures" as much as those who fail after graduation. Since organizations exercise different standards for retaining less amenable clients for treatment, even when their admission policies are the same, the use of the failure rate makes interorganizational comparisons more valid, particularly between public and private institutions. Further, a better understanding of treatment emphases may be gained by comparing internal and external failure rates across institutions.

It is rarely possible to show that innovative and progressive programs for delinquents improve their success in avoiding reinstitutionalization. However, they do not *increase* the failure rate. Therefore, such programs can be justified as more humane ways of dealing with youth.

Levine, Abraham S. "Evaluating Program Effectiveness and Efficiency," *Welfare in Review*, vol. 5, No. 2 (1967), pp. 1–11.

Making a cost benefit type analysis of social welfare programs is usually very difficult, primarily because many of the benefits of the program cannot meaningfully be translated into monetary terms. Before any measurement of intangibles is attempted, it is first necessary to determine what should be measured, and why. A distinction should be made between the out-

puts (services rendered to clients) and the benefits, and relevant change theories should be applied to the situation. Theory, in this case, can be of two types: content theory, which identifies relationships to be investigated, and methodological theory, which indicates how these relationships should be measured. In social work, theory is drawn from three main streams: psychotherapeutic, small groups, and organizational.

A general theoretical orientation is not enough to direct a good evaluation, however, and it is essential to develop an impact model specifying all the variables which should be subject to measurement, and their relationships to one another. Few projects exhibit all aspects of this comprehensive theoretical and methodological approach. Five on-going studies, each of which incorporates one or two of the aspects of the approach advocated by the author, are discussed.

Levine, Robert A. "Evaluating the War on Poverty" in *On Fighting Poverty: Perspectives from Experience*, edited by James L. Sundquist, pp. 188–216. New York: Basic Books, 1969.

This article reviews the evaluations being done concerning the War on Poverty. Criteria for evaluation and methods of measurement are needed at three levels: (1) Success of the War on Poverty as whole must be defined, and means of measuring costs and benefits must be found. (2) Criteria and measurements must be found for the evaluation of individual programs. (3) Criteria must be prepared to compare the effectiveness of individual programs.

Poverty is defined in terms of command over economic resources, which may be measured in several ways, e.g. number of people below the poverty line, or the number of dollars necessary to close the "income gap" of those below the poverty line. Only the first measurement has been used in evaluations of the poverty program as a whole, although it is less informative than the second. Measurements so far have failed to show that reduction in the number of poor is a direct result of the poverty program, rather than of other intervening factors, such as fluctuating employment levels.

For evaluating individual programs, two sets of criteria must be used: *proximate effects* (for

example, the effectiveness of a health program in improving the health of the poor) and *antipoverty effects*. These must be related to one another conceptually and causally. Four different categories of programs are identified: manpower programs, individual improvement programs, community betterment programs, and income-maintenance programs. Evaluation efforts in these four areas are critically reviewed.

The relationships among the various programs can be evaluated on three levels: (1) relations among programs within the four categories; (2) the relationships of the categories to one another; and (3) the relationships of major thrusts, such as service delivery versus social change. Little comprehensive work has been done in this area, but it is the most important for determining resource allocation to various programs and for theory building.

Levinson, Perry, "Evaluation of Social Welfare Programs: Two Research Models," *Welfare in Review*, vol. 4, No. 10 (1966), 5–12.

This article compares two models for evaluation of social welfare programs. The goal-model approach studies three kinds of variables: program variables, intermediate variables, and dependent variables in an effort to assess performance against explicitly stated goals. The basic flow through the goal-model begins with an incoming group ("income") possessing certain characteristics, to whom something is done ("input-output"), to achieve a desired change in intermediate and dependent variables ("outcome"). Program variables include the set of organized stimuli which are expected to effect change. One way to describe program variables is through an impact model, which is a set of theoretical concepts that trace the dynamics of how the program is expected to produce the desired effect. Such impact models are difficult to apply to new programs in a state of transition and without theoretical underpinning. Intermediate and dependent variables cannot be separated from each other, since we assume a causal link between them; in the case of work training programs, examples would be "level of employability" and actual successful employment.

The system model approach is based on a series of statements about the relationships among the various components of an organiza-

tion as it implements the program under study. It takes into account the fact that each program of the organization competes for scarce resources and that relations between the organization's programs and the external environment must be considered. Thus while an employment program under Title V might be a failure under the goal-model approach, it might be considered successful when its effects on staff morale or on the external community are taken into account or its results are compared to other organization programs.

Finally the author asserts that cost-benefit analyses are possible in evaluating welfare programs, for example by computing the degree of attitude change (intermediate variable) per X amount of money expended on the program.

Levitan, Sar, "Facts, Fancies, and Freeloaders in Evaluating Antipoverty Programs," *Poverty and Human Resources Abstracts*, vol. 4, No. 6 (1969), 13–16.

On the basis of a review of government projects and their evaluations (or lack thereof), it is concluded that critical evaluations are not being performed. This is, in part, a result of project directors' reluctance to support critical reviews of their efforts, and in part because it is safer for academics to publish speculations which are untestable than evaluations of ongoing welfare programs. For example, the government is currently spending $420 million on a program (JOBS) that it knows virtually nothing about. Yet there is little encouragement for a hard-headed review. Most of the money currently invested in government-sponsored research is yielding very few workable results. Until such concrete results are obtained, money will continue to be poured into vast, uncharted programs.

Likert, Rensis and Ronald Lippitt, "The Utilization of Social Science," in *Research Methods in the Behavioral Sciences*, edited by Leon Festinger and Daniel Katz, chapter 13. New York: Holt, Rinehart and Winston, 1953.

The readiness of non-social scientists to use social science results is dependent on three sources of motivation: problem sensitivity, an "image of potentiality" which implies that conditions may be changed to be better and more effective, and a general experimental attitude

toward innovation. Even where these pre-conditions exist, it is essential to set up good channels of communication between the researchers and the users. Conferences and meetings may help to provide the motivation and insight needed to put findings into effect.

Research methods are discussed in reference to research in formal organizations, but many of the comments are generalizable to other areas. Topics covered include: (1) Creating a cooperative atmosphere, avoiding resistance, and creating realistic expectations of what the research can do. (2) Whether organizational insiders or a research staff from outside should be employed. It is concluded that on the whole, outsiders will usually be more effective than insiders. (3) When self-surveys should be used. (4) The necessity of establishing responsibility to administrators at a level above that which is being investigated, in order to protect the research process and its integrity. (5) The problem of identifying basic rather than superficial variables. These must be diagnosed by the researchers, but they must also be accepted by the practitioners if the research is to be effective. (6) Obtaining a balance between theory and applied objectives. If there is too little theory, the results of the study may not be generalizable, and the researcher may find himself swamped in minutia. Furthermore, since applied settings are constantly changing, if there is no theoretical basis to the study, the results may be totally irrelevant to the needs of the client. (7) The need for confidentiality, if full cooperation is to be insured among the line employees. (8) Preparation of staff at lower levels to accept research results. This may involve participation in planning and interpretation by relevant personnel (which will also serve to enrich the study). (9) The need for quick presentation of preliminary findings which will help to maintain interest. (10) Participation of the researcher in organizational self-analysis when results are presented. Resistances have to be recognized and worked through. Timing and pacing of presentations—letting the organization assimilate evidence at its own pace—are important, as well as presenting the results in a positive atmosphere. Arbitrary insistence on the correctness of the data is inadvisable, and the results should be in a simple, nontechnical language. Analysis of

data presentation in a form that presses for action may facilitate utilization. Illustrative cases are presented.

Lippitt, Ronald, "The Use of Social Research to Improve Social Practice," *American Journal of Orthopsychiatry*, vol. 35, No. 3 (1965), 663–669.

Six patterns of use of scientific resources are identified (1) derivation of action designs from relevant research findings, (2) the adoption of experimentally tested models, (3) diffusion of ideas among practitioners, (4) diagnostic evaluations by outside researchers to feed back information, (5) self-study within the organization or community, supervised by outside researchers, (6) the development of collaboration between the consumer and scientist. Unfortunately, all too often the consumer has received no training in how to use the services of the scientist. Such training is necessary.

Six differences between social science utilization and the use of physical and biological science are identified.

(1) Most significant adoptions of new educational or social practice require significant changes in the values, attitudes, and skills of the practitioner. This requires deeper personal involvement, and there will be more resistance.

(2) Most innovations are *adaptations*, not adoptions like the use of a new drug. The adapter must be oriented toward the basic principles of the innovation if the adaptation is to be creative.

(3) The concept of "social invention" has not been adequately developed. Description of innovations is therefore poor.

(4) The practitioner gets very little feedback about the effectiveness of the innovation. It is easier to evaluate effects in the physical sciences.

(5) Practitioners in mental health and education are relatively isolated from their peers. There is little competition and incentive to innovate. There is also more fear of public reaction.

(6) The resources and networks to link applied and basic fields have not been developed. *Linking* agents are needed for consultation and training.

Lippitt, Ronald, Jeanne Watson, and Bruce Westley, *The Dynamics of Planned Change,*

pp. 263–272. New York: Harcourt, Brace and Company, 1958.

In this section of the book, methodological problems of research on planned change are discussed. When the change is in psychological and sociological areas, there are difficulties in obtaining any kind of valid measurement. Many of the variables are internal-process phenomena (such as improvement in mental health) without clear external symptoms. Furthermore, measures of change which are produced may not reflect actual subjective experience. The problem is to find the best way to measure aspects of a given situation, without losing sight of the larger complex of variables in which the factors of interest are buried.

Change does not follow a single course, nor a regular pace. When the change occurs in individuals, measurement is complicated by the fact that people do not all start off at the same baseline, and it is difficult to measure the relative amount and quality of change which each undergoes. Change occurs in spurts; it is clear that something is happening in the "latent" periods between rapid changes, but we know little about it. Sometimes a system must even move backwards before it can move forwards, which provides a further complication. Other dilemmas include interpretation of baseline measurements, determining the possible differences between self-selected populations and others which are apparently similar, and separating the effects of planned change from "natural" change.

Difficulties may also occur in the relationship between the research team and the change agent team, although each has much to gain from the other in insight and knowledge. Compromises must be made in order to break down defensive barriers which may result from lack of appreciation of the other's perspective. Another source of tension is competition for the time and attention of the client system. Even when the research and intervention are performed by the same person, problems may arise in attempting to reconcile the demands of the two roles. The best protection against conflict is a unified plan which has been worked out cooperatively in advance.

Lipton, Douglas, Robert Martinson, and Judith Wilks, *Treatment Evaluation Survey* (tenta-

tive title). State of New York, (forthcoming 1970).

This book is an analytical survey of all studies of correctional treatment published from 1945 to December of 1967 which met minimal research criteria. It includes an extensive annotated bibliography and a critique of post-adjudicatory treatment studies.

Longood, Robert and Arnold Simmel, "Organizational Resistance to Innovation Suggested by Research." Paper presented at the 57th Annual Meeting of the American Sociological Association, Washington, D.C., August 30, 1962. (*Unpublished*)

The thesis of this article is that organizations have built-in biases against adopting the conclusions of social science researchers. The three main reasons are individual personalities, the organization itself, and the culture in which the organization is embedded.

Although social scientists consider themselves to be objective, the cultural values of our society may blind them to certain conclusions in their research. Findings which are too disruptive of cultural values may also be rejected by the sponsoring organization. A second problem in getting organizations to adopt recommendations is that organizations, and the people who populate them, have stakes in perpetuating the status quo. Innovations pose a threat to the continued stable existence of organizations and are therefore likely to be adopted only with reluctance.

Personality factors and hostilities may be either random or systematic. In the case of public health organizations, antagonisms to social scientists tend to be more widespread than would be expected by chance. The authors conclude that social scientists are themselves partly to blame. By segregating themselves from other disciplines, by refusing to come to grips with the practical problems of the public health field, and by not sharing responsibility for the operation of the organization, they may give the impression that they have little of value to contribute to the field. If they participate in the rough and tumble of organizational decision-making, there is a better likelihood that their research results will be used.

Luchterhand, Elmer, "Research and the Dilemmas in Developing Social Programs," in *The*

Uses of Sociology, edited by Paul F. Lazarsfeld, William H. Sewell, and Harold L. Wilensky, pp. 513–517. New York: Basic Books, 1967.

Problems which frequently arise in the researcher-practitioner relationship in an action-oriented agency are listed and discussed. The research staff may face impossible demands for "instant knowledge" from administrative heads, or requests to do a variety of chores inappropriate to their position. For the researcher who is inexperienced in dealing with bureaucracies, such demands may result in strains which destroy the action-and-research relationship.

Administrators face the problem of whether "insiders" or "outsiders" should conduct the research. This is seen as a false dilemma by the author, who notes that objective role performance is more a matter of professional identification than of organizational membership. The danger of assimilation, or loss of autonomy, is the most crucial problem. Researchers tend to be naive about power, and are ineffective in dealing with non-academic power contexts. In this context, the action staff is the "establishment" and the researchers are the "minority." Administrative concern with negative findings may result in pressure to suppress them, or in changing goals in the middle of the project. Protection against such actions is only possible if the action-researcher relationship is well developed initially.

A final problem arises when the administrators have overly high expectations about research results. Administrators want firm, clear generalizations which may be applied to particular clients or situations, but research findings are reported in terms of probabilities which apply to specified populations only. This tends to alienate practitioners and reinforce the myth that research "gets in the way" of action programs. If the social scientist is not aware of the delicacy of his relations with action staff, the usefulness of demonstration projects may become negligible.

McCord, William and Joan McCord, *Origins of Crime: A New Evaluation of the Cambridge-Somerville Youth Study.* New York: Columbia University Press, 1969.

This study is a reexamination of the Cambridge-Somerville Youth Study, a project begun in the 1930's which attempted to prevent delinquency in boys by means of friendly, intensive counseling. The study took 325 matched pairs of 11-year-old boys in Cambridge and Somerville, half of whom were judged pre-delinquent and half normal. They were randomly assigned to control and treatment groups. Counseling and friendship continued with some boys from 1939 to 1945. Others dropped out because of change in neighborhood, military service, and death over the years.

Several evaluations of the project have been done. Forty percent of both treatment and control groups had some kind of criminal record by 1948. The present study, begun in 1955, followed the records of the Cambridge-Somerville boys, primarily to check the effectiveness of treatment ten years after its termination.

Court records were used as the most practical measure of the program's success in preventing criminality. Lifetime court records were secured for each boy in the study. The voluminous case histories for each boy were analyzed. Then from the files on childhood data and official court records, criminals were compared to non-criminals. Early evaluations had found no differences between treatment and non-treatment groups in number and types of crimes committed. This study found a tendency toward decreasing criminality with age, but this was not a function of the treatment. In short, the general approach of the project failed, but intensive therapy seems to have succeeded with a few boys.

The origins of crime are analyzed and several factors relating to the genesis of crime are discussed. (1) Intelligence was not strongly related to causation of crime. (2) Physical condition did not affect incidence of crime. (3) Social factors were not strongly related to criminality. (4) Home atmosphere had a very strong effect, with quarrelsome, neglective homes most conducive to crime. (5) Consistent discipline tended to prevent criminality. (6) Paternal absence, cruelty, or neglect tended to produce criminality. (7) The role model of the father was significantly related to criminality. (8) Mother's personality was the most fundamental influence in the genesis of criminality. Loving mothers had very few criminal sons. (9) Son's

position in the family structure and attitude had an effect on crime.

McDill, Edward L., Mary S. McDill, and J. Timothy Sprehe, *Strategies for Success in Compensatory Education: An Appraisal of Evaluation Research.* Baltimore: The Johns Hopkins Press, 1969.

In reviewing evaluation research material on compensatory education, the authors address themselves to three problems: (1) What can be done to improve the quality of evaluation research? (2) Where research is of acceptable quality, does it provide us with the needed knowledge about the success of the program? and (3) What lessons can be learned concerning effective compensatory education and its measurement?

Three different programs and their evaluations are compared in detail, and eight more in brief. Measurement problems outside of the control of the evaluator are discussed. Program factors which appear to be closely associated with success are listed, and two successful programs which met these criteria are described. The criteria are: (1) careful planning and clear statement of objectives, (2) small groups and individualized instruction, (3) materials closely linked to program objectives, (4) intensity of treatment, (5) teacher training in appropriate methods. The dilemma of evaluation is that of maximizing program flexibility while at the same time maximizing knowledge about what is effecting the change. The authors urge that some controlled and some flexible evaluation be done, while most should compromise between the two extremes.

McIntyre, Robert B. and Calvin C. Nelson, "Empirical Evaluation of Instructional Materials," *Educational Technology,* vol. 9, No. 2 (1969) 24–27.

Evaluation of instructional materials has usually meant the review of materials by experts, who judge their quality and general suitability. Further field evaluation is necessary, identifying the results of the applications of material in actual situations, as opposed to judgments based on assumptions about the etiology of learning and the material's own characteristics. To make such evaluations requires careful statement of the educational objectives to be achieved with each set of materials, and the degree of teacher involvement and competency required for successful use. Evaluations should then produce a statement of the probability of success and efficiency.

Main, Earl D., "A Nationwide Evaluation of MDTA Institutional Job Training," *Journal of Human Resources,* vol. 3, No. 2 (1968), 159–170.

The purpose of this evaluation was to determine whether the MDTA vocational training program had any effect on the income and employment of trainees during the period between the program and the interview. Data for the study were derived from interviews held early in 1966 with 1200 former MDTA and 1060 "controls" who were unemployed about the time the training courses started. The control group was selected through a partial matching process. A snowball sample technique was used, in which each trainee selected was asked for the names and addresses of friends who were unemployed. When these sources were not fruitful, a matched individual was obtained by canvassing the block where the trainee lived. Despite such careful (although non-random) selection procedures, effort was taken to control for background differences through multiple regression analyses when comparisons were made. Baseline data collections were made before the training program. Drop-outs were compared with those who had completed the program. Results showed that the MDTA program had no effect on income among those who had a full time job after the program, but more of the "graduates" and the program dropouts were employed. Neither length nor type of training had a significant effect on full-time employment.

Mangum, Garth L. "Evaluating Manpower Programs," *Monthly Labor Review,* vol. 91, No. 2 (1968), 21–22.

This report is concerned with the Manpower Development and Training Act, the Vocational Education Act of 1963, the Vocational Rehabilitation Program, and the U.S. Employment Service. From the experimental period between 1961 and 1967, ten services are identified which have proved useful in lowering obstacles to employment and increasing job retention of the disadvantaged. The programs also have serious

shortcomings, but there are not adequate data for evaluation of strengths and weaknesses, and no program currently has a reporting system capable of producing such data.

Mann, John, "The Outcome of Evaluative Research," in *Changing Human Behavior*, pp. 191–214. New York: Charles Scribner's Sons, 1965.

In an attempt to make generalizations about effective strategies for changing behavior, the author examines the conclusions of evaluative studies in the fields of psychotherapy, counseling, human-relations training, and education. He selected 181 studies with a greater degree of social significance and a lesser degree of experimental error than most research. He analyzes the research designs used, the number of subjects and practitioners included, the nature of the sample, the setting in which the method was tested, the nature of the method itself, the change criteria employed, the types of methodological error in the evaluation, and the findings obtained.

The information provides answers to two central questions. First, what are the general characteristics of evaluative research, regardless of content? Second, what are the differences in character and outcome of studies that evaluate different program strategies? The conclusions are that most evaluative research uses the simplest possible experimental design, often crude and only partially satisfactory. The findings of the research are unrelated to the ways in which change is measured. Oddly enough, all types of measuring instruments demonstrated the existence of positive change with approximately the same frequency, 45% of the time.

Moreover, there were no significant differences when program content was considered. In spite of differences in program concepts, training procedures for practitioners, populations, and social conditions, evaluation research conducted in different areas is similar in character and outcome. The only clear finding is that change is demonstrated in approximately 45% of the studies.

The conclusion drawn is that evaluative research has failed. Programs are too complex to be evaluated under operating conditions. In place of evaluation the author suggests laboratory research for the study of behavior change

strategies. Only with tight controls, isolation of specific program components, and factorial analysis can generalizations be built up.

Marris, Peter and Martin Rein, "Research," in *Dilemmas of Social Reform*, pp. 191–207. New York: Atherton Press, 1969.

The book discusses the community programs funded by the Ford Foundation and the President's Committee on Juvenile Delinquency between 1960–64. The chapter on "Research" examines the conflicts that arose between the experimental perspective, which sought impartially to evaluate the effectiveness of the new social programs, and the action perspective, which sought adaptively to explore possibilities and exploit promising avenues.

The authors conclude that the interests of experimental research and action are not the same. Research requires a clear and constant purpose, which defines the choice of activities; consistent procedures; and no revision in the program until the sequence of steps and the evaluation are completed. Action concentrates on immediate next steps and it changes direction as events proceed. Systematic experimentation and social action cannot both be carried out in the same operation. Because action is pragmatic and flexible, it needs to be retrospectively interpreted; outcomes cannot simply be related to initial aims and methods, because these have undergone continual revision.

The evaluations of the community projects discussed in the book were constrained by the pressures of action. They could not sustain their commitment to the logic of the experimental method. The authors imply that they would have been well advised to abandon controlled experimental evaluation for exploratory social analysis of the wider program process.

Mauldin, W. Parker and John A. Ross, "Family Planning Experiments: A Review of Design," *Proceedings of the Social Statistics Section*, American Statistical Association, 1966, 278–282.

The article reviews four evaluations of family planning programs, the India-Harvard-Ludhiana population study, the Singur, India Study (an outgrowth of the previous study), the Koyang Experiment in South Korea, and the Dacca, East Pakistan Study. Others are

mentioned briefly and conclusions are drawn about the design of family planning evaluations.

The ultimate objective is to affect fertility, although intermediate objectives are often given. Program inputs are a composite of contraceptive methods, informational content, type of media, frequency of stimulus, and intensity of stimulus. Differences between control and experimental populations are used as measures of effectiveness. The criteria of effectiveness are usually acceptances and continued use, fertility levels and change, and surveys designed to give information about knowledges, attitudes, and practices of family planning. Units assigned to experimental groups vary widely but they are always areas, not individuals. There are many possible sources of contamination but most are very difficult to control, for the early studies have been of varying quality, but some have been designed very well. Now very sophisticated studies are needed to answer questions about long-range effects, differences in effectiveness, and side effects of the more promising devices.

Merton, Robert K. "Role of the Intellectual in Public Bureaucracy" in *Social Theory and Social Structure*, (1964 edition), pp. 207–234. Glencoe, Illinois: The Free Press, 1964.

The relationship of the social scientist to public policy is colored by the fact that his findings, as compared to those of the physical sciences, are often indeterminate. This tends to undermine the relations which exist between experts and clients because of the difficulty of judging the expert's competence. It also increases the need of policy makers to rely on the judgments of experts in recruiting new expert personnel, which leads to the establishment of self-contained cliques. Further, policy maker may feel that experience has given him a considerable degree of competency in the intellectual's area of expertise, which makes the expert's role appear to be dispensable. Finally because the expert's investigations are concerned with alternatives which have value implications, he is especially vulnerable to attack by those whose sentiments and interests are violated by his findings.

Two types of intellectuals are identified: those who exercise advisory and technical func-

tions within a bureaucracy and those who are not attached to a bureaucracy. There is a danger that the intellectual in the bureaucracy may be converted into a mere technician. Specificity of demands and the exclusion of the social scientist from early stages of program planning may limit his role to one of gathering information. This is usually not true for unattached intellectuals, whose perspectives are less subject to the bureaucrat's immediate control. The problem for unattached intellectuals is rather to gain access to responsible policy makers. The crucial point is that the choice and definition of problems will be fixed in part by the intellectual's position within the social structure. The bureaucratic intellectual who must permit the policy maker to define the scope of his research problem is implicitly lending his skills and knowledge to the preservation of a given type of institutional arrangement. The unattached intellectual may bring forward knowledge which questions the existing system, even though he may have little effect on actual policy formation. These pressures on the intellectual in the bureaucracy may cause a change in his orientation: he becomes "less theoretical and more practical," and learns to think in terms of implementing policies within a given situation. Even when such accomodations takes place, there may still be conflicts in values between the intellectual and the businessman, as well as cleavages resulting from different social and power positions and general mistrust of the other's life styles and motivations.

The high turnover of expert personnel in public bureaucracies is therefore not merely a matter of client dissatisfaction; it is also often the product of the cumulative frustrations experienced by the intellectual.

Meyer, Alan S. and Stanley K. Bigman, "Contextual Considerations in Evaluating Narcotic Addiction Control Programs," *Proceedings of the Social Statistics Section*, American Statistical Association, (1968) pp. 175–180.

This paper discusses eight questions raised by evaluation and suggests ways in which they should be approached in the evaluation of narcotic addiction control programs. (1) What is the sponsor's purpose in having an evaluation? Evaluation in the addiction field may fulfill a

variety of perceived needs: to see how well the program is doing, to document the need to maintain the program, to find out how to improve the program, to stimulate fundamental innovations, to delay action in a controversial area, or many others. The evaluator should determine the real purpose in order to decide whether an evaluation will be meaningful or relevant. (2) What are the goals? Addiction control program goals tend to be unclear, contradictory, and shifting. This is largely a result of our ambivalent view of addicts as simultaneously criminal and ill, the extremely high failure rate and subsequent staff frustration, and the severe governmental constraint to which treatment programs are subject. (3) What is included and what can be excluded from the evaluation? If the evaluator can participate in answering these questions, he is more a behavioral scientist than a data analyst. (4) What should be the criteria for success? Drug programs have traditionally stressed abstinence as the primary criterion for success because of powerful political pressure. Recently social functioning and rehabilitation variables have been added. (5) How should success criteria be measured? Arrest records and employers' records have been used instead of supposedly unreliable addicts' self-reported behavior. To measure drug use, urinalysis has been traditionally used. There is controversy over the use of urin sis, however, since it is seen as degrading. (How should recipients be classified? Traditio classification systems have used personality v ables of participating addicts, but relevant soc classifications (eg. decreased criminality, i eased conventionality) are also essential fo helping programs to define individual problem) Who should do the evaluation? It is more important that the evaluator assume the role of behavioral scientist than whether he is an outsider or insider. (8) What are the constraints on dissemination of the findings? Any unjustified restraints on the distribution of results should be challenged.

Meyer, H. J. and E. F. Borgatta, *An Experiment in Mental Patient Rehabilitation.* New York: Russell Sage Foundation, 1959.

This study is of special interest because it illustrates the dependence of the evaluation team on the working of the program. In this case, the process by which clients for the program were referred and selected (a process over which the evaluation team had no control) resulted in a very small number of cases. This factor jeopardized both the validity and the reliability of the results of the study.

The evaluation concerned a shelter workshop operated by the Altro Health and Rehabilitation Services, which had recently agreed to accept psychiatric post-hospital patients. The program was intended to ease the transition from hospital to normal living, and reduce the likelihood of rehospitalization. The clients were engaged in manufacturing hospital uniforms, and were paid for what they produced; in addition, casework, vocational counseling, job placement, and educational assistance were available.

The workshop planned to take 80 patients during the two year study. However, extensive screening and the anticipation by the screening agents of the kinds of clients whom the workshop might favor resulted in the referral of only 41. Of these, 12 entered the program. The effect of the selection process was to include only those who were highly motivated or those unable to make adjustment outside of the hospital.

The design of the evaluation was experimental, and treatment and control groups were selected at random within the group of patients eligible to enter the program. The measure of recovery was avoidance of recommitment. Secondary criteria included inclusiveness of social relations, economic independence, reality of orientation, and self-attitudes. The results showed that the program was slightly advantageous.

Meyer, H. J., E. F. Borgatta and W. C. Jones, *Girls at Vocational High.* New York: Russell Sage Foundation, 1965.

The purpose of this study was to evaluate a program that identified potentially delinquent high school girls and involved them in preventive programs designed to reduce their expected rate of delinquency. Treatment goals were: increasing self-understanding, developing more adequate psychological and social functioning facilitating maturation, and supplementing inadequate emotional resources.

Girls entering a New York vocational high school between 1955 and 1958 were screened to identify those with potential problems. From

the pool of "latent deviates," a random sample of 189 girls was referred to the Youth Consultation Service, where they were offered casework or group counseling services by social workers, and 192 girls became the control group (Neither the school nor the agency was given the names of the controls.) Pre-tests and post-tests were administered to both groups at the beginning and the end of the school year.

Measurement tools used included general personality inventories, projective tests, self-reports of social adjustment and attitudes, and sociometric questionnaires. Measurements of school performance and behavior (suspension, expulsions, dropouts, truancy, attendance and conduct), and out-of-school behavior (out-of-wedlock pregnancies, known delinquent acts, and contact with police) were also made. None of the measures revealed significant differences between treatment and control groups.

Miller, Delbert C. *Handbook of Research Design and Social Measurement.* New York: David McKay Company, 1964.

This book is a compilation of resources designed to provide references to the essentials of research design. It includes general guides to research design and sampling, a guide to statistical analysis, descriptions of selected sociometric scales and indexes, and a guide to research costing and reporting.

Miller, Richard I. *Evaluation and "PACE": A Study of Procedures and Effectiveness of Evaluation Studies in Approved PACE Projects, with Recommendations for Improvement.* Report No. 1 of the Second National Study of PACE. Fairfax County Public Schools, Virginia, Center for Effecting Educational Change, February 1968. (*Unpublished*)

This evaluation of evaluations analyzes 21 proposals to evaluate the Project to Advance Creativity in Education. The proposals were analyzed: (1) to see if they met the requirements of the Stufflebeam model, and included provision for all four classes of evaluation—context, input, process, and product, (2) to see if the proposed evaluation procedures met minimal criteria of validity, reliability, etc., (3) to classify proposals with reference to design, means of data collection, population to be sampled, criteria of interpretation, and agents involved in the planning and execution of the evaluation. It was found that almost all the proposals lacked adequate theories, models, and designs, that the personnel were not trained, and that data collection techniques and processing facilities were inappropriate or insufficient. The authors recommend that new agencies should be established to fill these lacks.

Miller, S. M. "The Study of Man: Evaluating Action Programs," *Transaction,* Vol. 2 (March/April 1965), pp. 38–39.

In order to carry through effective social action projects, planners need sophisticated and reliable intelligence reports which can only be provided by evaluation research. But the project director may not know what information he needs, and unless he is highly involved and competent, he may not know how to use the information which is provided. The researchers' main problem, however, is the complexity of assessing the effects of community action programs. Different communities and programs will require different data and methodologies. Cooperation between researcher and administrator is essential.

Miller, Walter B. "The Impact of a 'Total-Community' Delinquency Control Project," *Social Problems,* vol. 10, No. 2 (1962), 168–191.

The delinquency control project discussed was conducted in a lower-class section of Boston between 1954–57. The evaluation concerns the effect of intensive therapeutic work with seven neighborhood gangs. A social worker was assigned to each gang, and after achieving rapport with the members, attempted to modify their behavior through the provision of an adult middle-class role model, and the introduction and encouragement of legitimate activities. The age range of the gangs was 12–21, although each gang was limited to a few years in the age-span. Four of the gangs were white-male; one was black-male; one white-female and one black-female. The total population comprised 205 individuals.

One specific goal of the behavior-modification program, the inhibition and control of delinquent activities, was studied. Trends in disapproved forms of customary behavior (swearing,

drinking, etc.), illegal behavior, and rate of court appearance were analyzed. Trends within each gang and across the target population as a whole were studied by comparing both self-reported and observed behavior at the beginning and the end of the project periods. Disapproved behavior was significantly reduced in only one gang (white, male, younger, higher social status) and significantly reduced for the whole group only in one of 14 behavior areas ("school-oriented" behavior). Illegal activity was reduced in both female gangs, but showed an increase, particularly in major crimes (11.2%), for males.

Trends in court appearances were measured using a quasi-experimental, time series design. First, objective court records over a 12-year period were gathered for each boy. Second, data on a control group consisting of matched gangs from the same neighborhood which had not had assigned social workers were also obtained. Within the target population, there was no significant decrease in the number of court appearances during the contact period. When the control group was compared with the target group, behavioral patterns appear exactly the same, and the decrease in court appearances after the contact period is therefore attributable to maturation rather than the effect of the social workers.

The author concludes that the evaluation shows that the delinquency control project had no effect on the law-violating or morally-disapproved activities of the gangs.

Morehead, Mildred, "The Medical Audit as an Operational Tool," *American Journal of Public Health*, vol. 57, No. 9 (1967), pp. 1643–1656.

This article discusses the use of the medical audit (a judgment of the professional performance of a physician by a clinician-surveyor) in determining the quality of medical care being given to a group of patients. The Health Insurance Plan, consisting of 32 medical groups, organized a study (1) to assess the performance of preventive health measures, (2) to assess the management of 10 cases of specified illness, and (3) to discover administrative and professional relationships, for each of the more than 400 family physicians. Surveyors, who had both teaching appointments

and clinical practices, rated the physicians on their records, diagnostic management, and treatment and follow-up for each case. For physicians who were below average in quality of medical care, corrective measures were suggested. The standards for new family physicians in the Plan were raised on the basis of characteristics found to be related to outstanding physicians.

Another study of costs and quality of medical care was undertaken by the Teamsters to assess their hospitalization plan. The study was similar to the HIP study. The Teamsters were able to use the results to inform their membership of the components of good medical care. The two studies were used to strengthen codes for surgery and other specialties. The administrative usefulness of this technique for discovering weak areas in medical care is great. These studies show that even the subjective quality of a physician's competence with patients can be measured with appropriate evaluation techniques.

Moss, L. "The Evaluation of Fundamental Education," *International Social Science Bulletin*, vol. 7, No. 3 (1955), 398–417.

Fundamental education is a process whereby people of low economic and social development are helped to consciously change their living conditions by their own efforts. The purpose of evaluation is to assist the operation of fundamental education programs by helping to achieve operational efficiency. Because such evaluation is different from academic research, different methods are appropriate. They should be limited and simple in operation. The results of the evaluation can thus be comprehensible to those who are affected by it. As the people develop, it becomes desirable for them to do their own self-evaluation in the process of self-organization and self-understanding.

All problems cannot be solved by limited techniques, but examples are given to show that a contribution can be made by relatively simple methods. Fairly direct methods, applied locally, may be used to measure three fundamental components of levels of living—nutrition, housing, and health. Opinion research methods can obtain measures of achievement and satisfaction. Exhibitions and films may be used to

encourage the development of a desire for change.

Unless the people are able to use the lessons of evaluation, the evaluation loses much of its value. Thus, the methods used in evaluating fundamental education should be capable of producing results which can help the people to objectively consider their own work.

Nagpaul, Hans. "The Development of Social Research in an Ad Hoc Community Welfare Organization," *Journal of Human Relations*, vol. 14, No. 4 (1966), pp. 620–633.

The potential conflicts between the researcher and the practitioner are documented in a case study of a government-sponsored community organization. A separate research branch was set up to do "action research" and "evaluation research" which would help heads of service units develop further program innovations. But the relationships between the research branch and the program units were not adequately formalized and there was no structure to resolve conflicts. The service staff resented the demands which the research staff made on their time, and many viewed the research program as a waste of money. The failure of the research staff to produce significant results—applicable feedback or basic research—led to deemphasis on the research program. Although some of the factors which led to the failure of the research program were organization-specific, many of them exist within other service organizations. The author concludes that it is time to reexamine the "demonstration project" concept in order to determine how it can more successfully combine the two professed aims of service and stimulation of experimentation.

Ott, Jack M. "Classification System for Decision Situations: An Aid to Educational Planning and Evaluation," *Educational Technology*, vol. 9, No. 2 (1969), 20–23.

This article describes the role of the Ohio State University Evaluation Center in a collaborative effort with the Ohio public school system. The Evaluation Center saw its function as providing information for decision making, and felt that it should be involved at all levels of the project. The role of the evaluator is to anticipate decision situations and problems

which may arise, and to provide relevant information which would otherwise be unavailable due to lack of foresight.

Owens, Thomas R. "Suggested Tasks and Roles of Evaluation Specialists in Education," *Educational Technology*, vol. 8, No. 22 (1968), 4–10.

The public wants evaluations of the overall effectiveness of educational programs, but educators also need more information on other issues relevant to decision making, such as the strengths and weaknesses of various programs, how they might be improved, etc. This necessitates evaluation of goals, plans, and operational procedures, as well as effectiveness. Reasons for the poor state of the art include a shortage of evaluation specialists and educational facilities to train them, a lack of an adequate evaluation theory, and poorly developed strategies of design and methodology. Another major problem is that the roles of the evaluator have never been adequately defined, and the various tasks which he may perform have not been enumerated.

Eleven tasks are presented and discussed: (1) developing a climate among practitioners that is supportive of evaluation, (2) planning and focusing an evaluation, (3) selecting or constructing instruments, (4) collecting data, (5) processing data, (6) analyzing and interpreting information, (7) reporting information, (8) assisting decision-makers in utilizing evaluation information, (9) providing support services not directly related to evaluation—such as helping to uncover implicit assumptions and real goals, or providing information on similar projects to the one being evaluated, (10) performing new research related to evaluation—such as testing evaluation models, and (11) administering an evaluation team. Because of the diversity of functions which an evaluator may be expected to fulfill, the author proposes that specialist roles be developed in evaluation. Such roles would include a director of research, a coordinator of the research project with other departments in the project, a surveillance specialist to bring new ideas back to the project and the evaluation unit, and specialists in the various stages of the evaluation process—instrumentation, data collection, data processing, and data reporting.

Parsell, Alfred P. "Dynamic Evaluation: The Systems Approach to Action Research," SP–2423 (Systems Development Corporation, Santa Monica). Paper presented at the American Sociological Society, September 1, 1966. (*Unpublished*)

The evaluation problems which are associated with large-scale, complex, diffuse community action programs (typified by those sponsored by OEO) are discussed. Most of these programs become involved in multiple interventions in an open environment. As such, standard research designs are very difficult to apply meaningfully. If not developed within the framework of the total operational context, such studies have limited value.

"Dynamic evaluation," the author's term for his conceptualization of evaluation, is an effort to place action-research in the theoretical framework of systems science, and thereby provide it with logical justification, an effective set of procedures, and an organization for the collection, analysis, and feedback of relevant data and findings. In use, the systems approach in "dynamic evaluation" views social action programs as action systems and regards information as a necessary concomitant to effective action system functioning. The kind of information required, together with its sources, uses, and flow are defined by the system itself, which presupposes careful definition and analysis of that system and its components. The dynamic evaluation approach endeavors to comprehend the total action system as a continuous and changing process, in which direction and control of the change are central. Systematic use of this process may be expected to lead to greater theoretical relevance and sophistication in problem-solving and to more efficient and effective action.

Perry, Stewart E. and Lyman C. Wynne, "Role Conflict, Role Definition, and Social Change in a Clinical Research Organization," *Social Forces*, vol. 38, No. 1 (1959), 62–65.

This case study documents the conflicts which occur when individuals (in this case clinical therapists) attempt to play the role both of practitioner and researcher. Such conflicts lead to individual redefinitions of roles in two ways. (1) The individual reintegrates his role by giving priority to one or the other of these conflicting roles. This may involve either a formal discussion between the therapist and the patient, or an informal recognition by the therapist-practitioner on his own. (2) A split-relationship definition may be worked out between several individuals. This involves role segregation, one individual interacting with a patient solely as a therapist, and another solely as a researcher. The difficulties in giving equal weight to both roles is illustrative of the inherent conflicts which may arise in "in-house" evaluations.

Plutchik, R., S. R. Platman and R. R. Fieve, "Three Alternatives to the Double-Blind," *Archives of General Psychiatry*, vol. 20 (1969), 428–432.

Classical drug-testing design uses the double-blind situation, where neither doctor nor patient knows whether the treatment or a placebo is being received. Three alternative approaches to this method of controlling bias in drug studies are discussed: (1) attempting to measure the extent of bias, (2) attempting to distribute the bias equally over all conditions or groups, and (3) exaggerating sources of bias in order to determine "some maximum combined effect."

Provus, Malcolm, "Evaluation of Ongoing Programs in the Public School System," in *Educational Evaluation: New Roles, New Means*, edited by Ralph W. Tyler. The 68th Yearbook of The National Society for the Study of Education, pp. 242–283. Chicago: NSSE, 1969.

This paper is a discussion of the Pittsburgh Evaluation Model. The purpose of program evaluation is to determine whether to improve, maintain, or terminate a program. Evaluation is the process of agreeing on program standards, determining whether a discrepancy exists between some aspect of the program and the standards governing that aspect of the program, and using discrepancy information to identify weaknesses of the program. The process of evaluation consists of moving through stages and content categories in such a way as to facilitate a comparison of standards and performance, while at the same time identifying standards to be used for future comparisons. Four stages are defined: definition, installation, process, and product. Evaluation procedure will be based on raising relevant questions in each stage. Each of the questions imply a criterion, new information, and a decision.

A major tenet of the Pittsburgh model is that each of these steps should be handled by different people. The formulation of the question is the job of the evaluator, the criterion is the responsibility of the program manager, and the collection of information to be used is a function of both evaluation and program-staff activity. The decision alternatives are outlined by the evaluator while the choice between alternatives belong to the program director. The activities of the program-staff and the evaluation-staff will mesh with each other in a series of linked question-information steps. The feedback of discrepancy information which may be useful to the staff will often mean that the first stages of the model will have to be re-appraised at several points, since new information will result in changes in the program.

The four stages are discussed in detail, and pertinent questions and likely problems are presented. Several important points are made about the assumptions on which this model is based. (1) It assumes that the evaluation team will probably not have a chance to participate in the planning of the program, since most educational programs are installed before the arrival of the evaluation team. (2) Important evaluations can be made in early stages of the program, even though final effects are not yet visible. (3) Evaluation and program staffs will be able to cooperate easily. (4) Efficient operation of program activity is dependent on effective evaluation activity. (5) Until the program is in the final stages of development, there is no need for experimental evaluation design. Experimental design applied too early may stifle improvements in the program and cause it to be rejected for the wrong reasons.

The author concludes that those involved in public school evaluations must recognize (1) the natural developmental stages of any new program, (2) the evaluation activity that is appropriate to each stage, and (3) the dependence of administrators on information obtained through evaluation if they are to make sound *defensible decisions*.

Regal, J. M. *Oakland's Partnership for Change.* Oakland, California: Department of Human Resources, June 1967. (*Unpublished*)

This report summarizes the history of the Ford Foundation's Gray Area Project in Oakland. It describes the programs funded by Ford and by agencies of the Federal government (particularly OEO), and presents evaluative data on their effectiveness. As the Research Advisory Committee to the project notes in an introduction, each evaluation study was conducted on an individual ad hoc basis (just as each project was conducted as a separate undertaking); no effort was made to evaluate the over-all effectiveness of the Oakland program. The Advisory Committee recommends that in future efforts, the over-all strategy of the Oakland effort be made explicit and that evaluation research should be directed at testing the effectiveness of the individual projects in accomplishing the strategic ends.

Regal notes that Oakland was unique among Gray Areas Projects in its insistence upon rigorous evaluation. Problems arose, such as resistance by program staff to completing forms and to using systematic approaches in accepting clients for service. Programs as operating were different from the programs described in applications for funding. The evaluator was rarely successful in convincing program managers to return the original design. Nevertheless, important results emerged.

Regal suggests that further evaluation should test assumptions which program managers treat as truths, such as that decentralization brings services to people who otherwise would not receive them, that indigenous workers have greater rapport with ghetto residents than professionals and can build a more helpful relationship, that compensatory education programs can help close the gap between poor and middle-class children.

Riecken, Henry W. "Memorandum on Program Evaluation." Prepared as a staff paper for the Ford Foundation, October 1953. (*Unpublished*)

Four types of evaluation are identified—*effect studies*, where the focus is on goal achievement; *operations analysis*, which reports on activities and the extent of program implementation; *surveys of need*, which assess the situation and the potential worth of the

program; and *investigations,* which may involve punitive action. Four general categories of programs suitable for evaluation are presented. These categories are based on rough similarities in subject-matter, "size," flexibility, degree of the evaluator's control over the situation, and data gathering techniques. The categories are: information and education programs, skill or performance training programs, microcosmic social welfare programs, and macrocosmic social welfare programs. Two factors not included in this typology should also be used to distinguish types of programs: degree of institutionalization, and manifest and latent objectives. The author presents a list of 11 suggestions for improving the state of the art in evaluation. These range from preparing annotated bibliographies of existing evaluations to the creation of institutes and specialized training programs for evaluators.

The second section of the paper discusses some major technical and procedural problems in evaluation research. (1) A six-step model for evaluation studies is developed. Each step is defined in some detail, and a comment is made on problems and constraints which may be encountered. The steps are determining program objectives, describing operations, measuring effects, establishing a baseline, controlling extraneous factors, and detecting unanticipated consequences. (2) The relationship of evaluation to action is examined. The evaluation must be built into the program from the beginning, since at least some effort to obtain baseline information should be made. The evaluator must also make sure that his plans are backed by the highest authority in the agency. The choice between the "objective outside evaluator" versus the "familiar inside evaluator" is discussed. (3) In choosing problems to study, the evaluator should try to optimize conditions conducive to technically competent research, a list of which is offered. He might choose to stay away from settings where few of these conditions are met. (4) If evaluation research is to become a cumulative discipline, more attention should be given to social variables rather than to assessment of overall programs.

Robinson, John P., Jerrold G. Rusk, and Kendra B. Head, *Measures of Political Attitudes.* Ann Arbor, Michigan: Survey Research Center, 1968.

This volume lists scales for measuring political attitudes that are useful for survey research. After a historical summary of public opinion, the book presents measures in ten areas: liberalism-conservatism, democratic principles, domestic government policies, racial and ethnic attitudes, international affairs, hostility-related national attitudes, community-based political attitudes, political information, political participation, and attitudes toward the political process.

Specific items, scoring instructions, data from research studies using the items, and assessments of the strengths and limitations of each scale are also provided.

Robinson, John P. and Phillip R. Shaver, *Measures of Social Psychological Attitudes.* Ann Arbor, Mich.: Survey Research Center, 1969.

This volume lists empirical instruments for measuring a number of important attitude areas, the actual items in the scales along with scoring instructions, and an assessment of their strengths and weaknesses. The areas covered are: life satisfaction and happiness; self-esteem; alienation and anomia; authoritarianism, dogmatism, and related measures; other socio-political attitudes; values; general attitudes toward people; religious attitudes; methodological scales.

The authors also present a review of current knowledge in each domain, data from research studies using the listed attitude scales, and references to the literature.

Rodman, Hyman and Ralph L. Kolodny, "Organizational Strains in the Researcher-Practitioner Relationship," in *Applied Sociology: Opportunities and Problems,* edited by Alvin Gouldner and S. M. Miller, pp. 93–113. New York: The Free Press, 1965.

The nature of the role of the researcher within a professional agency and the difficulties which stem from this role are the focus of the paper.

Often personality factors are blamed for strain between researchers and practitioners. Personalities are not irrelevant but they often serve to mask the nature of the role relation-

ship. Organizationally-structured stresses are often overlooked. The question of credit for publication, for example, is a chronic problem in the relationship. Because of the evaluative nature of the researcher's role and his special tie to the administrator, the practitioner often feels that the researcher has a vested interest in discerning and reporting errors. The researcher's primary job is tangential to the practitioner's primary job, and they organize their time differently, thus making it harder for them to understand each other or to collaborate effectively. The researcher is in a marginal position in relation to the practitioner and this may intensify his ties to the administrator, adding to the strain. Denial and displacement by the practitioner, the development of one-way communications, and various changes in the formal organization are manifestations of the strain. Although some strain is inevitable, it is possible to alleviate it through recognition of its primary source, the social organization of the clinical agency.

Rosenblatt, Aaron, "The Practitioner's Use and Evaluation of Research," *Social Work,* vol. 13, No. 1 (1968), 53–59.

In order to determine the value of research for the social work practitioner, a study was done with four groups of social workers. The purpose of the study was disguised. Data were collected concerning the respondent's use of research in handling difficult cases, the value of research findings for his practice, the helpfulness of research in improving his practice, and the helpfulness of research courses in preparing the respondent for his career. The findings were viewed from four perspectives: (1) the extent to which workers read research articles in treating a difficult case against other alternatives like consulting supervisors or colleagues, (2) workers' ratings of the value of supervision, consultation, and research, (3) workers' ratings of the helpfulness of various experiences (practice pre- and post-graduation, supervision, inservice training, reading) in improving practice, (4) workers' opinions about the helpfulness of research as opposed to other courses in preparation for their career.

In each area, research was rated the least used or least helpful activity. Possible explanations are the type of person who goes into social

work, the irrelevance of research findings to practical problems in the field, lack of agreement between researchers and practitioners on definitions of basic concepts, etc.

As professionals, social workers are committed to improving their practice. As part of that commitment, they must support research, cooperate with researchers, and pay attention to research findings. Only then can the available knowledge be organized and put to maximal use.

Rossi, Peter H. "Boobytraps and Pitfalls in the Evaluation of Social Action Programs," *Proceedings of the Social Statistics Section,* American Statistical Association (1966), pp. 127–132.

New treatments for social ills are unlikely to produce massive results in a highly developed country. For example, the introduction of universal education produced a large-scale impact, but the introduction of further refinements (such as Head Start) are likely to have only marginal results. The discrepancy between expectations for the program and results have often caused evaluation findings to be rejected.

The controlled experiment is the best, most precise means of evaluating programs. Because of the difficulty in obtaining control groups, the design of "placebo" treatments is suggested. Often, however, it is not possible to conduct even this type of experimental evaluation, and only comparative or quasi-experimental designs are possible. Such "soft" techniques are adequate for the detection of massive effects, and if they produce a result of "no significant difference," it is unlikely that more precise techniques would show more than marginal differences. It is therefore suggested that evaluations be carried out in two stages: a reconnaissance stage, where "soft" correlational methods are used to screen out programs which appear most worthwhile, and then an experimental phase, in which differential effectiveness may be precisely tested.

Sadofsky, Stanley, "Utilization of Evaluation Results: Feedback into the Action Program," in *Learning in Action,* edited by June L. Shnelzer, pp. 22–36. Washington, D.C.: Government Printing Office, 1966.

This paper discusses evaluation research conducted for the purpose of feedback into the project being evaluated. It reviews some of the obstacles to utilization of feedback. Three relevant questions are discussed. (1) Do program operators or other decision-makers accept evaluation findings as credible? (2) Since evaluation results are only one of a number of inputs into the decision-making process, do decision-makers feel that change in the action program is warranted or necessary? (3) Are changes possible? In other words, are other alternatives likely to achieve the desired objectives, and if so can they be implemented? Discussion of these questions indicates the wide variety of factors that can obstruct utilization of evaluation results.

Scanlon, R. G. "Innovation Dissemination," *Pennsylvania School Journal,* vol. 116 (March 1968), 375-376.

This article reports on some of the criteria found to be necessary for the successful *implementation* of a newly developed educational program, viz. IPI (individually prescribed instruction). These criteria include administrative commitment in each school; teacher commitment; administrative and teacher retraining for new roles; administrative and teacher participation in research concerning the program's evaluation; and consideration to individual local conditions in reference to the school's readiness for the program.

Schulberg, Herbert and Frank Baker, "Program Evaluation Models and the Implementation of Research Findings," *American Journal of Public Health,* vol. 58, No. 7 (1968), 1248-1255.

Evaluation research findings are seldom implemented by program administrators. This paper attempts to determine ways to enhance the implementation of evaluation results. Only when the purpose of administrators is really to use the results of the evaluation are findings likely to be implemented. The evaluation must be based on criteria meaningful to funders and administrators.

There are two basic approaches to evaluation, the goal attainment model and the system model. The goal attainment model, which measures the success or failure of a program in reaching certain defined objectives, is characterized in practice by a lack of concern with implementing results. If the researcher accepts the program administrator's goals as his criteria, he often finds that stated organizational goals are not the goals of the real-world program, and thus his evaluation is meaningless. Since real goals are interrelated, the evaluation of one or even several specific goals is artificial in the context of a complex organization. An evaluation of one or two specifics is very difficult to implement, given the multitude of constraints imposed by other factors.

The system model, which attempts to determine how closely the organization's allocation of resources approaches an optimum distribution, is more likely to result in implementation. It is concerned not only with goals but also with such other functions as coordination of sub-units, resource allocation, and adaptation to the environment. It uses its data to suggest intra-organizational and organizational-environment linkages and feedback mechanisms which bridge the gap between evaluation findings and program modification. Feedback can be enhanced by evaluation procedures which fit the decision-making needs of an organization and by making data available when needed. The system model involves more expensive and time-consuming research, but results are more readily usable. Organizations should establish planning divisions to ensure the translation of research results into program planning.

Scriven, Michael, "The Methodology of Evaluation" in *Perspectives of Curriculum Evaluation,* edited by R. W. Tyler, R. M. Gagné, and M. Scriven, pp. 39-83. Chicago: Rand McNally, 1967.

The focus of this paper is on curriculum evaluation, although many of the points are easily transferrable to other types of evaluation research. Particular stress is laid on the deficiencies in present evaluation practice and means for reducing these deficiencies.

The first part of the article deals primarily with defining the evaluation study. A distinction is made between two functions: the *formative* function involves on-going appraisal while the program is being constructed and tested, while the *summative* function is concerned with making a decision about the success of the program

CONTINUED

1 OF 2

after it has been developed and run. A task such as curriculum development must include evaluation from its inception, since those in charge of creating the program must constantly be evaluating their progress. The author warns, however, that the intrusion of the professional evaluator (employing rigorous standards of measurement) at too early a stage in a program may stifle creativity.

A second group of topics deal with some practical and theoretical problems which arise in evaluating goals. Although many evaluators feel that it is most important to examine how well a program has achieved its stated goals, Scriven argues that it is also necessary to evaluate the worthwhileness of the goals themselves. Two major methods of evaluation are "intrinsic" and "payoff." The "intrinsic" approach raises many problems because it brings in intermediate goals (such as the elegance or integrity of the program) as well as final goals. "Payoff" evaluations, on the other hand, are concerned only with the end performance of the program, and may therefore be deficient in explaining the whys and processes of success or failure. The author suggests that way of optimizing the desirable qualities of both is through "mediated" evaluation, which involves continuous reassessment of the fit between goals and the program, as well as a final test of the results.

Another set of issues is whether or not comparative research is desirable in the case of curriculum evaluations, and if so how it can be carried through in a situation where the "double-blind" is impossible. The author gives a detailed example of how a variety of new cheap curricula can be developed as controls in evaluating a new "super curriculum", in order to control for interfering factors such as teacher enthusiasm. Finally, criteria are presented for assessing educational achievement and other critical variables which might be affected by curriculum change.

Sharp, Laure M. and Rebecca Krasnegor, *The Use of Follow-Up Studies in the Evaluation of Vocational Education*. Washington, D.C.: Bureau of Social Science Research, 1966. (*Unpublished*)

This report investigates the availability of follow-up data on students of vocational education and the utility of the data for evaluating the effectiveness of vocational education programs. It then offers recommendations for improving the information base both substantively and methodologically.

Follow-up studies are available in some geographic areas and for certain types of programs. Where available, they are useful tools for evaluating vocational training. Future studies should remedy the neglect of certain regions (e.g. the South) and programs (e.g. those in technical institutes and junior colleges), and pay greater attention to (1) the employment situation, (2) the labor market, and (3) characteristics of schools (type, size, role of teachers, curriculum content, guidance and placement services, etc.). Longitudinal nation-wide control-group studies should be supplemented by intensive small-scale studies of particular programs and issues.

Sharp, Laure M., Barton Sensenig, 3rd, and Lenore Reid, *Study of NDEA Title IV Fellowship Program: Phase I*. Washington, D.C.: Bureau of Social Science Research, March 1968. (*Unpublished*)

Phase I of the evaluation was based on secondary analysis of records of the first four years of the NDEA Fellowship Program. Comparisons were made between NDEA fellowship recipients and other doctoral candidates for whom data were available from the National Academy of Sciences Register of Earned Doctorates and an NORC survey of 1961 college graduates.

Thirty per cent of NDEA grantees in 1959–1962 had completed their doctorates by June 1966. Fellows who completed their doctorates did so somewhat more rapidly than a "similar" group of non-NDEA fellows. Over half of NDEA Fellows with doctorates were engaged in teaching. Within each academic area, NDEA Fellows were more likely to go into teaching than non-NDEA doctoral recipients. Since the program was designed to increase the number of college teachers, the findings indicate some measure of success. No data were available on the educational careers of Fellows who had not obtained the doctorate. Phase II will collect additional data.

Shaw, Marvin E. and Jack M. Wright, *Scales for the Measurement of Attitudes.* New York: McGraw-Hill, 1967.

This volume is a compilation of some 200 attitudinal measures in eight substantive areas: social practices, social problems, international issues, abstract concepts, political and religious problems, ethnic and national groups, significant others (including the self) and social institutions. Descriptions of the measures, types of subjects on which they were developed and tested, and procedures for scoring them, and data on reliability and validity are given.

Sheldon, Alan, "An Evaluation of Psychiatric After-care," *British Journal of Psychiatry,* vol. 110 (1964), 662–667.

This study focuses on the issue of whether after-care of discharged mental patients can in fact prevent re-admission to any significant extent. The study sample was drawn from patients discharged from Warlingham Park Hospital, Croydon, England, between October 1961 and March 1962. Only women between the ages of 20 and 59 with a diagnosis of schizophrenia or depression were included. A total of 89 successive patients fulfilling the criteria were randomly allocated either to psychiatric after-care (45) or referred to their general practitioner (44). Within the after-care group, patients were randomly allocated either to a day center or an out-patient department. The day center patients' primary therapeutic relationship was with a nurse, while those in the out-patient clinic were seen by a doctor. Patients were followed for six months and post-test data were obtained on 83 cases. For after-care patients, the six-month readmission rate was 18%, compared with 47% for patients referred to general practitioners. There was no difference between the day center and out-patient clinic cases. Psychiatric after-care was associated with a longer time spent under care in the follow-up period, but shorter subsequent hospitalization. Good attendance, inversely associated with re-admission in all three groups, may be more important than type of after-care.

Sheldon, Eleanor B. and Howard E. Freeman, "Notes on Social Indicators: Promises and Potential," *Policy Sciences,* vol. 1 (1970), 97–111.

The development of social indicators has received a great deal of public and professional attention, but there is little consensus on what types of indicators are most relevant. This vagueness has encouraged confusion about the potential utility of social statistics for planning, program development, and scholarly endeavors. Three "impossible" uses of indicators are discussed. First, social indicators will not help to make policy decisions more objective, because the very choice of important indicators assumes a value orientation and a set of priorities. Statistics can be used by any advocate to argue for his position. Second, indicators may not be used as a satisfactory substitute for evaluation research, since it is impossible at present to control for all potentially interfering variables. Third, the use of indicators for social accounting is meaningless unless there is a theory which defines all the important variables and their interrelationships. Economics has such a theory, but the other spheres of social science do not. On the positive side, however, the development of social indicators may lead to improved descriptive reporting of trends, and therefore to better analysis of social change. Increased understanding of past social changes may in turn lead to more effective prediction of future social events. At present, those interested in change find themselves "data poor" and collection of social statistics would help to remedy this situation.

Sherwood, Clarence C. "Issues in Measuring Results of Action Programs," *Welfare in Review,* vol. 5, No. 7 (1967), 13–18.

The problems that plague evaluation are both technical and sociopolitical. The latter are more important and intransigent. Evaluation strategy should encompass all the means for getting and maintaining support for the evaluation effort and for creating a climate for the incorporation of findings into the decision-making process.

To overcome resistances to spending scarce funds on evaluation, ways must be devised to make the most efficient use of evaluation dollars and to show society the benefits of evaluation. Simultaneous evaluation of many programs with the same objective is one way to improve evaluation efficiency and utility.

More effort should be devoted to spelling out the behavioral "variables that are crucial to the solution of social problems" (e.g. character, problem solving ability) and to devising valid measures of them. More research should be done on evaluation research itself, such as surveys of community understanding of evaluation and its purposes, further development of research instruments and models, and training of all people associated with programs in the concepts and procedures of evaluation.

Silver, George, *Family Medical Care.* Cambridge: Harvard University Press, 1963.

The Family Health Maintenance Demonstration aimed at combining prevention and treatment of both physical and emotional disorders, emphasizing health promotional activities. Family health was seen in terms of appropriate functioning in work, play, sex, and family life. The demonstration was carried out for eight years in the context of a prepaid medical group affiliated with HIP and the Montefiore Hospital. Treatment teams of internists, social workers and public health nurses gave preventive and therapeutic services and consulted with medical specialists. In addition to regular comprehensive medical care from the hospital medical group, study families received family guidance and emotional support from the health teams for up to four years.

Families were assigned to study and control groups. Demonstration teams collected initial, interval, and final data on study families and examined controls at completion. Twelve areas of family function were examined. Physicians rated family medical history and each member's physical condition. Nurses rated nutrition, sleep, and rest, educational achievement of children, recreational adjustment, and housing. Social workers evaluated personal adjustment, family relationships, and occupational adjustment of father. Each area was rated excellent, good, fair, or poor on a scale from 4 to 1. Members of study families improved in physical condition from 2.7 to 3.1, but were not superior to the controls at the end of the study (3.0). At final evaluation 88.5% of study subjects rated good or excellent vs. 75.6% of controls. Improvement was noted in 4 of 5 areas evaluated by nurses, especially housing. Of the 5 areas evaluated by the social worker, the average rating declined in 4. There was no greater improvement among study families than among controls in educational achievement of children. Controls did better in 9 out of 12 evaluation areas. Improvement in physical condition was not related to high utilization of medical services or to prevalence or absence of symptoms and difficulties.

Editor's note: The negative findings of this demonstration project evaluation indicate that self-serving biases are not inevitable in internal evaluation.

Sjoberg, Gideon and Roger Nett, *A Methodology for Social Research.* New York: Harper and Row, 1968.

The purpose of this book is to examine the logical and theoretical bases of methodology rather than specific techniques and procedures. The authors proceed from the perspective of the sociology of knowledge. Of special interest is the emphasis placed throughout the book on the relationship of the researcher to research. The scientist is part of a larger social system, within which the research subsystem exists. The assumptions, norms, and values of these social systems impinge on every phase of the research process, from the selection of a topic to the analysis of data. Furthermore, the stages of research are highly interrelated: initial conceptualization and design will determine, or at least set limits to, the type of analysis which may be employed. Considerable attention is given to the logic of inquiry, the development of theoretical systems, and the connections between theoretical and methodological systems. On a more concrete level, contrasts are drawn between what is done in the research process and what should be done.

Slocum, W. L. "Sociological Research for Action Agencies: Some Guides and Hazards," *Rural Sociology,* vol. 21, No. 2 (1956), 196–199.

Sociologists who undertake applied research in action agencies should have personal, human-relations competence as well as academic qualifications. They should also preferably have some experience in the area being studied, so that they will understand the agencies' problems. It is important to build confidence in the research among the agency heads, and to establish a working, two-way channel of communication,

where the researcher makes an attempt to translate his ideas into non-technical language. Agency opinion leaders should be encouraged to participate in the identification of problem areas in order to help establish a working relationship, and they should also be involved in the translation of research findings into action proposals.

Before the social scientist decides to undertake research for action agencies, he must weigh the evidence as to potential hazards: mis-use of research (to delay action, or to defend an entrenched position), administrative interference with the research process, lack of cooperation in data collection, and problems of publication should the findings be negative.

Smith, Bruce L. R. *The Rand Corporation*, pp. 216–237. Cambridge: Harvard University Press, 1966.

Research conducted by individuals within an organization tends to be bland, raises no disruptive problems, and rarely reflects unfavorably on the organization. Research done by outsiders has the advantage of access to top decision-makers. An outsider role also dissociates the study from intra-agency conflict and makes it difficult to discredit the study on the basis of the interests, institutional or personal, of any special group. The Rand Corporation is a model of an outside research agency that can serve as an institutionalized critic, sufficiently independent to be provocative, yet closely related to the points of decision within the organization. This type of relationship will probably gain increasing acceptance in the future.

Smith, Joel, Francis M. Sim, and Robert C. Bealer, "Client Structure and the Research Process" in *Human Organizational Research*, edited by Richard N. Adams and Jack J. Preiss, chapter 4. Homewood, Illinois: The Dorsey Press, 1960.

This case study points up the problems of the independent social scientist who does research in a large bureaucratic organization. In this case, a liaison group was established within the organization, which represented all interested units. The representative character of the group made it difficult for them to operate frankly in negotiations with the researcher, and

also denied the researcher direct access to other people (and information). Second, although the researchers wanted a fairly loosely worded contract, pressure was put on them before the research had begun to specify methodologies and instruments. Third, bureaucratic personnel changed during the research period, and the successors had little understanding of informal agreements which had been reached. Thus, the formal contract became central.

Although the original idea had been to do an exploratory methodologically oriented analysis, the researchers found themselves tied down to the methodological approaches which had been specified in the contract, and were also pressured to produce quantitative, immediate problem-solving data. Stereotyped research traditions in the bureaucracy limited approaches suggested by the researchers: for example, they proposed to sample specific populations, but the department insisted that the sample be random. The interview schedule was monitored, and important questions were eliminated because the organization felt them to be sensitive. Only a single pre-test of the field instruments was allowed, which hampered methodological experimentation. The author concludes that although bureaucratic characteristics heightened the degree of interference with the research, the problems stemmed, in part, from the fact that the evaluative nature of the research was perceived as potentially damaging by some members of the organization and therefore was not consistently supported.

Somers, Gerald G. *Evaluation of Work Experience and Training of Older Workers*. Madison, Wisconsin: Industrial Relations Research Institute, University of Wisconsin, 1967. (*Unpublished*)

The author examines the adequacy of available data to do a cost-benefit analysis of work training programs for older workers who are receiving public assistance. After review of the data, he concludes that they are incomplete, particularly when the focus is on trainees aged 50 and over.

He identifies specific data gaps, such as the absence of earning and employment information prior to enrollment in the Work Experience and Training program and lack of two-year follow-up data. His recommendations include

the use of appropriate control groups, more and better information on the total population eligible for the program, cost data for each component of the program, more details on periods of employment and unemployment and income by characteristics of trainees, analysis of the causes and consequences of the large dropout group, weighting of non-economic costs and benefits.

Stake, Robert E. "The Countenance of Educational Evaluation," *Teachers College Record*, vol. 68, No. 7 (1967), 523–540.

Curriculum evaluators are expanding their role to include behavioral science variables as well as traditional tests, but the tendency is still toward a purely descriptive rather than judgmental presentation of results. Many evaluators prefer not to play the role of judge, but judgment is necessary. Evaluation should at least describe the merits and faults of the program as they are perceived by affected groups (teachers, parents, experts, etc.).

Three important categories of data are identified. *Antecedent* data deal with conditions existing prior to teaching of a program, *transactions* with the processes and events which occur in the program, and *outcomes* with program impact. Descriptive data (including the intent of the program and actual operations) and judgmental data (including standards used and actual judgments) should be gathered for all three categories. In analyzing descriptive data, the extent of congruency between intent and outcome should be noted, as well as the logical and empirical contingencies between antecedents, transactions, and outcomes.

In making judgments, the evaluators should describe the different standards which relevant groups may have and specify the criteria used by each. Judgment of merit may be made with respect to absolute criteria as reflected through personal judgment, or by comparison with other programs. Summative, final evaluations will be primarily concerned with judging by relative standards, whereas feedback, formative evaluations will be interested in charting contingencies and in using absolute standards to aid in program innovation and refinement.

Stake, Robert E. "Testing in the Evaluation of Curriculum Development," *Review of Edu-*

cational Research, vol. 38, No. 1 (1968), 77–84.

This article reviews and discusses models and methodologies, issues in curriculum evaluation and on-going evaluation studies. An extensive bibliography of good references is included.

Stein, Herman D., George M. Hougham, Serapio R. Zalba, "Assessing Social Agency Effectiveness: A Goal Model," *Welfare in Review*, vol. 6, No. 2 (1968), 13–18.

This article is concerned with evaluating an agency's effectiveness as an organization by seeking a methodology to measure the effectiveness of an agency in achieving its declared output goals. Output goals are the instrumental goals to change the allocation of resources into programs and their delivery to target populations. Output goals are the link between input goals—attaining adequate resources—and outcome goals—solving individual or societal problems.

A process analysis of the agency's flow of services would relate operations to the achievement of the agency's output goals. The output goals may be quantity of service, quality norms of service, or coverage of service in serving a defined population. The first step in assessment is identifying the agency's output goals, using the agency's articulated statement as the point of departure. Attempts to assess quality involve measurement of success rates in treatment or determination of the caliber of personnel employed. Relevant data concerning quantity goals, quality goals, and coverage goals must be accumulated. This goal model approach evaluates the agency as a delivery system by comparing its actual service output with its formal output goals.

Stephan, Frederick F. and Phillip J. McCarthy, *Sampling Opinions: An Analysis of Survey Procedure*. New York: John Wiley and Sons, 1958.

This book is geared primarily to the non-mathematical social scientist. Section one is an introduction to the variety and characteristics of sampling procedures, common sampling models, and general principles of sampling. Part two analyzes various common problems of sampling through the examination of actual empirical studies—comparison of survey estimates

with known population characteristics, comparison of estimates from several samples, estimation of variances, sampling variability of quota sampling procedures, accessibility and cooperation.

Part three deals with the design of sample surveys. Stress is placed on the preliminaries to actual design, including formulation of objectives, gathering of information on the population, assessing measurement tools in relation to the survey, and reviewing possible procedures. Design development topics cover the outline of the design, use of population subgroups, control of changes in original sample design during operations, and appraisal of survey performance.

Steward, M. A. "The Role and Function of Educational Research—I," *Educational Research*, vol. 9, No. 1 (1966), 3–6.

This article focuses on the relationship of the teacher to the researcher in a school system. Teachers are not afraid of research, but they would like to participate in establishing the problems to be researched, and they are annoyed by specialists who have little understanding of the teachers' problems. Teachers would also like to have research reports presented in language which they can understand. Researchers who feel that teachers are unable and unwilling to cooperate have not attempted to understand their perspective and their sense of exclusion from the research processes.

Teachers are beginning to have more contact with evaluative research and more understanding of its relevance. In larger school districts, research is becoming a recognized function, and there may be a permanent research officer on the staff. Because of the increase in research in schools, teachers colleges should attempt to familiarize their students with both the processes and findings of such inquiry. The dissemination of findings to practicing teachers is also a crucial step in involving the teacher in the cycle of research and utilization, and professional associations may play a large role in this area.

Stouffer, Samuel A. "Some Observations on Study Design," *American Journal of Sociology*, vol. 55, No. 4 (1950), 355–361.

Society rewards quick, plausible "answers", and tedious, modest experimental design is not

in demand. It is research of the latter type that provides the knowledge base for a cumulative science. Experimental designs are expensive and not always possible to conduct, but if the experimental model is kept in mind, research findings will be more valid and reliable. With forethought, it is often possible to obtain information which will approximate the missing elements of the "experiment."

An example is given of research on white soldiers' attitudes toward having Negroes in their platoons. Only "after" measurements were obtainable for the experimental and control group, but by asking the experimental group what their attitudes had been before the Negroes entered their platoon (and ascertaining the validity of these recollections), a tentative effort could be made to infer the effects of racial proximity on attitudes.

Even more important is choosing initial problems, or orientations, based in theory. Because experimentation is expensive, it is too great a luxury to conduct isolated fact-finding enterprises. When exploratory research is being done, limitation of the focus of the study to one or two variables at a time will result in more valuable contributions to theory than massive, ill-defined inquiries into highly complex phenomena.

Stromsdorfer, Ernst W. "Determinants of Economic Success in Retraining the Unemployed: The West Virginia Experience," *The Journal of Human Resources*, vol. 3, No. 2 (1968), 139–152.

This paper reviews the same project as Cain and Stromsdorfer's "An Economic Evaluation of Government Retraining Programs in West Virginia". (See Cain and Stromsdorfer). In addition to the findings presented there, there is discussion of the results of a multivariate analysis that isolated the most important sociodemographic and training variables.

Stufflebeam, Daniel L. "Evaluation as Enlightenment for Decision-Making." An Address delivered at the Working Conference on Assessment Theory, sponsored by the Commission on Assessment of Educational Outcomes, The Association for Supervision and Curriculum Development, Sarasota, Florida, January 19, 1968. Columbus, Ohio: The Evaluation

Center, College of Education, The Ohio State University. (*Unpublished*)

The poor state of the art of educational evaluation is traced to three causes—lack of trained evaluators, lack of appropriate instruments and procedures, and lack of adequate evaluation theory—of which the last is most important. The main conceptual problems which face evaluators are (1) poor understanding of decision processes and information requirements in current programs, (2) inadequate definitions of educational evaluation in relation to the emergent requirements for evaluation, and (3) lack of appropriate evaluation designs.

The purpose of evaluation is to provide relevant information to the decision-maker. Because of the wide scope of new projects funded by current federal legislation, there is a real need for a continuous cycle of evaluation. The author conceptualizes such a cycle as a series of interrelated information processing and decision making feedback loops at the local, state, and federal levels.

He proposes a clasification system for educational decision situations based on the functions which the decisions will serve. The categories of decision functions include *Planning* (goal specification), *Programming* (program specification), *Implementing* (program direction) and *Recycling* (major program modification). Given these four types of decision situations, there are four matching types of evaluations. *Context* evaluation will be used for program planning; *Input* evaluation will be used to help develop program activities, *Process* evaluation will be used to monitor the implementation of the project, and *Product* evaluation will assess results after a complete cycle of the program. Each of these types of evaluation should have a different methodology and conceptual framework which articulates with its objective. The logical structure of evaluation design is, however, the same for all studies.

Suchman, Edward A. "Action for What? A Critique of Evaluative Research," in *The Critique of Organization Management, and Tactics of Social Research*, edited by Richard O'Toole. Cambridge: Schenkman Publishing Company 1970.

Evaluation makes three asumptions (1) that man can change his social environment, (2)

that change is good, and (3) that change is measurable. An objective, a program, and criteria for measuring change, are essential. Evaluation must be related to a decision making process; if it is unlikely to be used, it should not be done. The evaluation should be timed right—after the program has become operational and, possibly, effective and before change is no longer possible. The idea of a self-contained, one-shot study of a clearly defined program is not appropriate in most real situations.

Demonstration projects are of three kinds: pilot programs, model programs, and prototype programs. Only with model programs is rigorous, controlled experimentation appropriate. For the others, flexibility and quick feedback are more important.

Once a program is in operation, the evaluation must focus on the improvement of services. This type of evaluation requires a model which stresses the feedback of a continuous stream of information into the ongoing process.

The basic design of evaluative research, whether of the before-after or during-during variety, must include a description and analysis of input, an understanding of the cause-effect process which leads to change, and a definition of the objective in terms which permit measurement of attainment. Three common designs are the case study, the survey of experimental and control groups after a program, and the prospective study done periodically on experimental and control groups.

Administrative and interpersonal problems of evaluative research vary with the type of evaluation undertaken, whether system-oriented or program-oriented. Conflict between program staff and evaluation staff is inevitable in any case. The best solution is to include the program staff wherever possible in developing the evaluation. This also increases the possibility that the evaluation will be utilized.

Suchman, Edward A. *Evaluative Research: Principles and Practice in Public Service and Social Action Programs*. New York: Russell Sage Foundation, 1967.

This excellent handbook for evaluation research draws primarily on the author's experience in the field of public health, but it is equally applicable to other fields. The book is divided into three main sections representing

the conceptual, methodological, and administrative aspects of evaluation. It begins with a brief historical account of evaluative research and a general critique of the current status of evaluation studies, with particular emphasis upon the shortcomings of many of the evaluation guides proposed for community self-surveys of public service programs. This is followed by a conceptual analysis of the evaluation process and an analysis of different levels of objectives and categories of evaluation.

The next section deals with methodology. A comparison between evaluative and non-evaluative research is made, and different approaches to evaluation are discussed. Research designs appropriate to evaluation are presented, and emphasis is laid on sampling procedures, isolation and control of the stimulus, and definition and measurement of the criteria of effect. Finally, reliability, validity, and differential results in the measurement of the program effects are discussed.

The administrative section includes chapters on the administrative process as related to planning, demonstration, and an analysis of resistance and barriers to evaluation; further discussion deals with problems in the administration of evaluation studies such as resources, role relationships, the carrying out of an evaluation study, and the utilization of findings. The book concludes with a brief exposition on the relationship of evaluative research to social experimentation, stressing the potential contribution which the study of public service and social action programs can make to our knowledge of administrative science and social change.

Suchman, Edward A. "A Model for Research and Evaluation on Rehabilitation," *Sociology and Rehabilitation*, edited by Marvin B. Sussman, pp. 52–70. Washington, D.C: American Sociological Association, 1966.

This paper presents two models for research: one for basic research on rehabilitation and one for applied evaluation. The evaluation model studies the relationship between a specific program and the attainment of some predetermined valued objective. The primary evaluation goal is to determine the extent to which an activity is associated with the occurrence of results, and the secondary goal is to test the validity of

the conclusion that the specific activity produced the effect. Only a valid base in knowledge and theory can give assurance that certain rehabilitation activities will achieve certain objectives. The inability to formulate evaluation hypotheses concerning which aspects of the rehabilitation process produce which specified desirable or undesirable results is due to a lack of basic knowledge and theory.

Defining clear program objectives is an area of difficulty in applied evaluation. General considerations involved in the formulation of objectives are: what are we trying to change? Who is the target? When is the desired change to take place? Is there one objective or several? What is the desired magnitude of effects? Most rehabilitation programs have multiple objectives resulting in a need for establishing an order of priority of these objectives.

Evaluation may be done as an assessment of effort, an assessment of effect, or an assessment of process. The evaluation may consist of an individual's or group's estimate of a program in which they are taking part, the appraised worth of an activity as given by a group of experts, or the scientific measurement of effectiveness made in term of acceptable standardized procedures.

There is a need in the area of rehabilitation for more scientific research which examines the objectives of a particular program (including its underlying assumptions), develops measurable criteria related to the objectives, and sets up controlled situations to determine the extent to which the objectives are achieved.

Sudman, Seymour, *Reducing the Cost of Surveys*. Chicago: Aldine Publishing Company, 1967.

This book, which is directed at professional researchers who are familiar with basic survey techniques, covers a wide variety of means to reduce the rising cost of large scale surveys. No single grand scheme is given, but each area of survey research methodology is examined separately for corners which may be cut, or new techniques which may be used without jeopardizing the quality of the research. Chapters include: probability sampling with quotas, the use of self-administered questionnaires, telephone interviewing, controlling interviewing

costs, and the use of computers and optical scanners.

Takishita, John Y. "Measuring the Effectiveness of a Family Planning Program: Taiwan's Experience," *Proceedings of the Social Statistics Section*, American Statistical Association, (1966), pp. 268–271.

This paper discusses the methods used in evaluating the intrauterine contraceptive family planning program in Taiwan, including fertility measures, insertion figures, and surveys. The most serious problem confronting the evaluators is the lack of knowledge as to what rate of fertility decline can reasonably be expected given the large number of factors which affect the birth rate.

Taylor, Philip H. "The Role and Function of Educational Research," *Educational Research*, vol. 9, No. 1 (1966), 11–15.

The *role* of educational research is to fulfill the needs of its consumers. Teachers want practical guidance, administrators want insights into the functioning of pupils and educational institutions, politicians want support for their positions, and parents want means to help their children achieve the social and personal goals which they have for them. Educational research has a responsibility to help fulfill all of these needs. The *goal* of educational research is to understand to what extent stated educational objectives have been achieved. A major problem is that some people do not really want the answer to this question, but would prefer a confirmation of their own beliefs. Others may desire objective information, but become confused in translating it into action. Because of prejudice and confusion, it is necessary to clarify the specific goals of research. Greater emphasis on methodological development is needed as well. An area of importance, which is highly related to utilization of research results, is the empirical study of decision-making processes in the school. This topic should receive greater attention in the future.

The field of education is full of people who want to judge; educational research should not be concerned with making judgments, but with presenting evidence. This attitude cannot be instilled in future researchers by telling them

about it; it must be demonstrated by those already in the field.

Therkildsen, Paul and Philip Reno, "Cost-Benefit Evaluation of the Bernalillo County Work Experience Project," *Welfare in Review*, vol. 6, No. 2 (1968), 1–12.

At the time of writing, the Title V Work Experience and Training project was in process. The paper describes the evaluative tools and methods that were developed for the evaluation, rather than results.

The evaluation is dealing with both *tangible* costs and benefits and *social and psychological* changes. One interesting feature is the development of an Employment Readiness Scale which measures progress from unemployment to successive levels of personal and family adjustment and of skills necessary to get and hold a job. Other measures being used include the Case Movement Scale of the Community Service Society, which is based on social worker judgments, the Semantic Differential to discern changes in self-evaluation by participants, and the Cantril and Free "Ladder Scale" to assess participants' expectations for the future.

The analytic design is expected to yield information on the sequence of steps through the program that provide optimal success, i.e. employment and self-support. This involves investigation of the selection of program participants; kinds of social and supportive services given; different training methods, skills, and length of training; type of job placement. Relationships between intangible changes and tangible dollar benefits and costs will be analyzed.

Thorndike, Robert L. and Elizabeth Hagen, *Measurement and Evaluation in Psychology and Education* (second edition). New York: John Wiley and Sons, 1961.

The uses of psychological and educational tests are dealt with in this book. Some attention is given to specific tests (an appendix lists many available tests) but the main emphasis is on when and how to use them. Chapters discuss: planning a test, qualities desired in any measurement procedure, where to find information about specific tests, standardized tests of intelligence, measurement of special aptitudes,

behavioral measures of personality, projective tests, planning a school testing program, etc.

Turvey, Ralph and A. R. Prest, "Cost-Benefit Analysis: A Survey," *Economic Journal*, vol. 75, No. 300 (1965), 683–735.

The general principles of cost-benefit analysis are discussed. These include (1) the enumeration of costs and benefits—a definition of the project and its scope and nature, costs to organizations or institutions other than the one sponsoring the project, secondary benefits which might accrue, and the length of the project; (2) the placing of dollar values on the costs and the benefits; (3) the choice of an interest rate to be used in the analysis; and (4) the determination of the relevant constraints on a cost-benefit analysis in the situation. Particular applications of the method are discussed, including examples from irrigation, transport, land usage, health, and education.

Tyler, Ralph W., Robert M. Gagné, and Michael Scriven, *Perspectives of Curriculum Evaluation* (AERA monograph series on curriculum evaluation, No. 1). Chicago: Rand McNally, 1967.

This monograph contains papers by each of the three authors, and a synopsis by J. Stanley Ahmann. The papers are: "Changing Concepts of Educational Evaluation," by Ralph W. Tyler "Curriculum Research and the Promotion of Learning," by Robert M. Gagné "The Methodology of Evaluation," by Michael Scriven (see abstract under Scriven).

Underhill, Ralph, *Methods in the Evaluation of Programs for Poor Youth*. Chicago: National Opinion Research Center, June 1968.

This report discusses a pilot study of poor youth conducted for the Office of Economic Opportunity. The aims of the study were (a) to determine the feasibility of longitudinal study of a national sample of poor youth and (b) to refine measurements of the situations and attitudes of teenagers and the correlates of their success, both of which were important for subsequent use in program evaluation.

The pilot study proved feasible; reinterviews were completed with 91 per cent of the youth. Usable indicators were devised on background, experience, ability, self-concept, and attitudes.

Factors associated with success in school and in the labor market were identified.

The author therefore recommends the use of longitudinal surveys of a national sample of poor youth as a "control" for evaluations of the effectiveness of antipoverty programs for this age group. He discusses techniques for assessing the extent to which the programs (rather than other factors) cause changes in the participants. Variables which are related both to selective entry into programs and to successful outcomes should be controlled. Multiple regression and demographic standardization are appropriate techniques. A good description of the relatively unfamiliar demographic standardization procedure is provided.

U. S. Bureau of the Budget, Executive Office of the President. *Household Survey Manual*, 1969. (Available from National Technical Information Service, U.S. Dept. of Commerce, Springfield, Va. 22151, document # PB 18 7444.)

The manual describes the concepts and definitions currently in use in Federal statistical agencies and the appropriate questions for measuring important basic characteristics of the population and the kind and quality of their homes. The "well tested and standardized" questions relate to items such as personal and family characteristics, employment history, education, income, health and disability, and housing. The manual furnishes advice on survey operation, including criteria for obtaining reliable samples. The last chapter proposes more tentative and subjective concepts and questions dealing with community characteristics and the attitudes of community residents. Extensive exhibits and appendices are included which give examples from previous national surveys, such as the census, school enrollment, work experience, work history, and housing occupancy and vacancy surveys.

U. S. Congress, House Committee on Government Operations, Research and Technical Programs Subcommittee, *The Use of Social Research in Federal Domestic Programs*, vol. III, 90th Congress, 1st session. Washington, D.C.: Government Printing Office, 1967.

Volume III of this four-volume study is entitled "The Relation of Private Social Scientists

to Federal Programs on National Social Problems." It presents responses of 53 eminent social scientists to an inquiry about the role of social scientists with regard to government policies and programs, and it reprints a collection of outstanding papers on the use of social research in policy-making. Authors of the papers include Robert K. Merton, Arthur Schlesinger, Jr., Daniel P. Moynihan, John W. Gardner, Rensis Likert, Marvin B. Sussman, Wilbur Schramm, among others.

Volume II is also worth the attention of evaluators. It focuses on social scientists assessments of federally financed research on six domestic social problems and agency assessments of the adequacy and quality of extramural social research.

U. S. Department of Agriculture, Agricultural Marketing Service, Food Distribution Division, *The Food Stamp Program: An Initial Evaluation of the Pilot Project*. Washington, D.C.: April 1962.

This is an example of a complex evaluation of a broad social program. The goals of the evaluation were to determine program feasibility, operating problems, effectiveness in increasing good diet among the eligible poor, and satisfaction with the program among relevant parties. The evaluation was carefully conceptualized, and evaluation procedures were built in before the program began, so that comparative measures could be gathered where relevant. Data were gathered not only from participants and administrators, but from relevant groups of grocers and social workers, whose cooperation was essential to the success of the program. Both formative and summative elements were considered. Although the main goal of the evaluation was to determine effectiveness, qualitative information about administrative problems and unanticipated consequences (such as changes in the sales patterns of retail stores) was also gathered, so that recommendations might be made about program changes if the data indicated that the program should be expanded.

An overall evaluation of administrative experience and problems was made, through interviews, in all eight demonstration areas. Four special evaluation studies were also con-

ducted. (1) A survey was taken of attitudes toward the program. Questionnaires were given to samples of moderate and higher income families in two of the pilot areas, participating and non-participating poor, food retailers, local welfare workers, and all state public welfare administrators. All groups exhibited positive attitudes toward the program, and welfare workers and participants-felt that it was more effective than surplus food distribution. (2) An analysis was made of the dollar volume of food sales in a sample of retail stores in the pilot area. Control data were gathered from stores serving middle and upper class individuals. Data were collected during a four week period to the start of the program, and a four week period after the program had been in effect for several months. After adjustment for seasonal factors, an 8% increase in grocery sales was shown. (3) Household food consumption surveys were conducted in two of the pilot project areas, before and after the inauguration of the program. An analysis was also made of matched households of non-participants. In both areas, participating families spent substantially more per person for food, and showed increases in the money value of all food consumed. Less increase was found in the rural area sampled, due to greater participation in a surplus food program previous to receiving food stamps. (4) A dietary evaluation was made of the families sampled in the household survey. Good diet was defined as receiving 100% or more of eight nutrients, as recommended by the National Research Council. Participating families had a considerably higher percentage with good diets than did non-participating families. No before measures appear to have been taken.

U. S. Department of Health, Education, and Welfare, *A Bibliographic Index of Evaluation in Mental Health*, prepared by James K. Dent. Washington, D.C.: Public Health Service, Publication No. 1545, October 1966.

An annotated bibliography of 300 items relating to the evaluation of mental health services is presented. The bibliography emphasizes the social aspects of evaluation, and includes a detailed index. The items cover nearly all the relevant literature for 1965, and nearly half the literature for the preceding ten years.

U. S. Department of Health, Education, and Welfare Office of Education, *Preparing Evaluation Reports: A Guide for Authors*. Washington, D.C.: Government Printing Office, 1970.

This guide is designed to help authors decide what to include in an evaluation report and how to report it. The recommended format is a brief summary of the objectives, methods, and results of the program, a section dealing with the setting or context of the program, a detailed description of the personnel, services, and procedures, analysis and discussion of the results of the program, and conclusions and recommendations based on the evaluation. Suggestions are made about what to include in each part. The guide includes questions which should be considered, followed by short explanations and examples. Sample narratives are given as simple models of parts of reports. References are listed, graded "easy," "harder," or "difficult."

U. S. Department of Health, Education, and Welfare, National Mental Health Advisory Committee, *Evaluation in Mental Health*. Washington, D.C.: Public Health Service, Publication No. 413, 1955.

The problems and processes of evaluation in mental health fields are discussed in this book, and an extensive annotated bibliography is presented. Chapter 1 deals with theoretical and methodological considerations, such as the difficulty of defining the concept of mental health, the bias of professional judgments, the anxiety created by scientific evaluation and the requirements for good research design. Chapter II presents seven general areas into which mental health evaluation studies may be grouped: community organization; administration; professional personnel; education and information; preventive effects of programs; factors influencing individual mental health; and diagnostic, prognostic, and treatment procedures. The scope of each of these areas is defined, and examples are given of the approaches (and in some cases the results) of studies in each area. Chapter III is a critical review of the present state (1955) of evaluative activity and a recommendation for higher priority for evaluation and for more effort at testing basic hypotheses about mental health.

Nine hundred eighty-four references are

listed and most of them are annotated; 107 deal with theoretical and methodological issues, and 877 are concerned with investigations in the seven mental health activity areas.

Vanecko, James J. *Community Action Program Goals for Institutional Change: Preliminary Report on National Evaluation of Urban Community Action Programs*. Chicago: Center for Urban Studies, University of Illinois, July 1969. (*Unpublished*)

Community Action Agencies were evaluated in order to determine which characteristics were associated with effectiveness in influencing other institutions to respond to the needs of the poor. The CAA projects were divided into three groups on the basis of distinct goal orientations: emphasis on education and social service, emphasis on employment, or emphasis on community organization. These three types of programs were then compared on 20 dimensions of change which occurred in the "target" institutional spheres. Dimensions included such items as increase in the number of people being served by social service agencies, increased agency efforts to hire members of minority groups, and increased participation by residents in school affairs.

By organizing the analysis on a comparative basis, the evaluation is able to provide more information about which CAA programs are most effective, and which institutional areas are most susceptible to change, than could an overall assessment of the impact of the program.

The results show that programs which emphasize educational and social service goals have little effect in producing institutional change; those that emphasize employment goals have a demonstrable effect in getting employers to hire their graduates; those that emphasize community organization have the most impact on changes in the public schools, neighborhood political life, and private social service agencies.

Vanecko, James J. "Community Mobilization and Institutional Change," *Social Science Quarterly*, vol. 50, No. 3 (1969), 609–630.

The preliminary findings of a national evaluation of Community Action Programs in 100 cities of 50,000 or more are reported. The evaluation included interviews with community

action agency board members, executive directors, local directors, related service agency members, PTA presidents, personnel directors, and neighborhood political leaders.

The effective CAP was found to have the following characteristics (1) central office support for community organization; (2) neighborhood centers actively involved in community organizing; (3) neighborhood centers uninvolved in militant activity; and (4) an absence of specific demands on other institutions. Effectiveness is highly related to two critical stages: gaining support of the board and executive director for an emphasis on community action, and developing community organization activities within the neighborhood center. (For greater methodological emphasis, see James J. Vanecko, *Community Action Goals for Institutional Change*.)

Wall, W. D. "The Future of Educational Research," *Educational Research*, vol. 10, No. 3 (1968), 163–169.

Teachers and administrators often find research too time-consuming, and they become impatient when apparently desirable programs are held up. As educational decision-making becomes more political, quick results are desired; this may result in impressionistic, hasty studies which are used to justify reforms which have not been sufficiently researched. This tendency may prejudice the development of sound research in education and the development of education itself.

The classroom teacher is confronted with choices, and research results must serve as a guide to choice, although other factors will enter into decision making. On all policy-making levels the final judgments about programs will be based on values—political, social, and moral. Research can define the limits within which choice is possible, and help to avoid confusion between what is desirable and what is practical. However, it cannot evaluate the worth of basic goals.

Evaluations cannot be done by the program developer, since he is already committed to the worth of his approach. Adequate evaluation is expensive, complex, lengthy and not to be undertaken lightly. Part of the reason for this is that evaluation tools are inadequate and clumsy: we are suffering from lack of past investment in

educational research and social science in general. *Action* studies must be expanded to provide information for decision-making, but this will take a long time. In the interim, limited *feasibility* studies must be done to identify problems of implementation and *evaluations* must be made to see if program aims are being achieved.

Wallace, David, *The Chemung County Research Demonstration with Dependent Multi-Problem Families*. New York: State Charities Aid Association, 1965.

This demonstration was intended to assess the effects of intensive professional casework with multi-problem families who account for a disproportionately large share of total welfare expenditures. Experienced graduate social workers of the Welfare Department and the Council of Community Services in Chemung County, N. Y., were assigned to the cases. They were given full professional freedom. Caseloads were limited to 20 for the supervisor and 30 for the caseworkers. The entire family was considered as the client and case closing was to be done on the basis of the satisfactory total functioning of the family. Median duration of treatment period was slightly less than 2 years. The research and treatment operations were carried out independent of each other. The study group met the following criteria: family was receiving Aid to Dependent Children or Home Relief, there was a mother figure and at least one child, the family received services from at least one agency 3 years or more before screening. 195 cases met these criteria. A control group was drawn from the same population as the demonstration families. There were 50 demonstration and 50 control cases drawn randomly from the pool. Then a third group of 50 cases was drawn to serve as a hidden control, known only to research staff. The demonstration group and first and control group were assessed before and after casework, while the second control group was interviewed only after the project. Demonstration and control group cases were rated on nine scales of social functioning (Geismar): family relationships and family unity, individual behavior and adjustment, care and training of children, social activities, economic practices, household practices, health conditions and practices, relationship to project worker, and use of community

resources. From research summaries of case records, three judge pairs rated families on a 7-point scale. The demonstration and control groups started at comparable levels. At closing the demonstration group had more cases at both positive and negative extremes. It was concluded that the procedures used in the study could not demonstrate a beneficial effect from whatever the caseworkers did.

Ward, David A. and Gene G. Kassebaum, "On Biting the Hand that Feeds: Some Implications of Sociological Evaluations of Correctional Effectiveness." Paper presented at the 61st Annual Meeting of the American Sociological Association, August, 1966. (*Unpublished*)

Prison departments and administrators are increasingly reluctant to sponsor evaluation research because of the fear of public reaction to negative findings. Even the most innovative administrators are becoming sensitive to the fiscal implications of evaluation findings that show little success. Correction professionals, to counter negative findings, often raise alternative criteria for judging the success of a program, e.g. improved emotional stability of inmates, rather than the "hard" criteria used by the evaluators, usually recidivism rate. In a 5-year study of counseling in a California prison with approximately 1,300 inmates, the criteria used were (1) the maintenance of order in the prison community and (2) the reduction of recidivism. The objective of the counseling program was a lessened endorsement by inmates of values which sanction anti-social behavior. The study included three treatment groups and a control group. Inmates were tested, interviewed, and observed; files were examined; group counselors were questioned, tested, and observed; parole agent reports and records were gathered. Of special importance were questionnaire data measuring inmate solidarity over time, reports of prison rule violations, and reports of arrests of parolees. A follow-up study was done over 2 years. There were no significant differences in outcome among the treatment groups as between the control group and the treatment groups. Endorsement of the inmate code was not lessened nor was the incidence of prison discipline problems.

The department of prisons wants to continue its present wide use of counseling and is therefore discounting the study and finding other outcome criteria to justify continuation. Prisons are less willing to allow outside scrutiny and are relying more on controllable intradepartmental evaluations. The current crop of correctional evaluations, which report no impact of treatment, may have mixed effects—either an increase in the variety of new programs, or restricted circulation of evaluation results. If results are not released, redundant studies will be done and ineffective programs may be maintained.

Wardrop, James L. "Generalizability of Program Evaluation: The Danger of Limits," *Educational Product Report*, vol. 2, No. 5 (1969), 41–42.

The two major types of evaluation, formative and summative, are mutually exclusive. The uncertainty principle indicates that one cannot simultaneously know *what it is* and *why it is*. Formative research on educational products will describe in great detail the specifics of the curriculum package, the environment in which it is tested, the reactions of the teachers and children, etc. This type of research on one case does not provide enough information to aid in decision-making about the adoption of the product in dissimilar environments at other times. It is not, like the summative type of research, designed for generalizability. Although the professional may easily be able to make this distinction, the consumer will make generalizations from results, even when such generalizations are unwarranted or illegitimate. Because the evaluator cannot control the use of his results, he has an obligation to design the study to have maximum generalizability, at least along the dimension of primary interest. This will necessitate the use of scientific as well as descriptive methods.

Webb, Eugene J., Donald T. Campbell, R. D. Schwartz and L. B. Sechrest, *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally, 1966.

Most social research relies on interviews and questionnaires to collect data. This engagingly written book urges more effort to use "non-reactive" measures, i.e. those not subject to biases from interviewer-respondent interaction

or other factors in the research setting. The authors describe a variety of imaginative measures that do not depend on asking questions, such as the use of records, observation, physical evidence, etc. They illustrate the use of these measures in specific research situations.

Weeks, H. Ashley, *Youthful Offenders at Highfields*. Ann Arbor: University of Michigan Press, 1958.

This book reports the results of an evaluation of a short-term, high-intensity, group therapy-oriented institution for delinquent boys. Three criteria were considered: (1) recidivism rates, (2) changes of attitudes, values, and opinions toward families, law and order, and outlook on life, and (3) changes in personality. The design of the study included pre-tests and post-tests on the attitudinal and psychological measures. A control group was selected from boys committed to the state reformatory, who would have been eligible for Highfields had there been room for them. Post-tests were administered immediately after release and again after the boy had been in the community for six months, in order to control for any "halo effect" which might be reflected in immediate post-release tests. Boys were followed up through their parole officers in order to obtain information about general adjustment and recidivism. An interesting aspect of the design was the interviewing of five of the boys' role partners (a parent, a friend, the proprietor of the place where he "hung out", a policeman, and his parole officer) just after his commitment and two months after he had been out, in order to determine whether they believed that there had been any real changes in the boy, and whether their own attitudes toward him had changed.

Interesting points emerging from the analysis include (1) an attempt to control for the somewhat different background characteristics of the two samples by constructing a prediction table, in which the combined effect of background and attitudinal factors were taken into account; (2) a comparison not only of success rates among "graduates", but among all entrants. Since Highfields had a higher internal failure rate, this represents a very important control; (3) the separation of the total groups into racial subgroups, since somewhat different

patterns emerged for each; (4) the finding that the attitude scales were not effective in measuring any changes in either the treatment group or the control group, although original attitudes were highly related to succes rates for both groups. (The scales were measuring changes in attitudes to which the program did not direct itself.) (5) The finding that there was no relationship between the psychoanalytic measures used, and recidivism rate. These latter two points indicate a somewhat inadequate theoretical basis in the program.

Results of the study show that recidivism among Highfields graduates is considerably less than among reformatory graduates, that the Highfields treatment did not appear to affect attitudes, but that it did effect some psychological rehabilitation-or at least retard further degeneration—as compared to the reformatory.

Weinberger, Martin, "Evaluating Educational Programs: Observations by a Market Researcher," *The Urban Review*, vol. 3, No. 4 (1969), 23–26.

This article makes several generalizations about educational research drawing on extensive experience with market research. To insure impartiality, evaluations should be performed by an agency that has no stake in the outcome. A climate should be created in which program failures are regarded as useful economies, not wasted efforts. The objectives of a program should not be evaluated by the researcher; they are matters of policy. The evaluator should be invited only to contribute to the conceptualization of the goals. In order for a program to be evaluated, the evaluator must know the objectives, the relative importance of each objective, and how much achievement of each objective makes a program a "success."

Evaluations should strive to pinpoint the causes of failure or success. Experimental design affords a means for isolating factors that contribute to failure or success. Magnification (which involves multiplying the effects of a small part of a project in order to estimate the effects of program expansion) is a risky but useful tool for learning about effects difficult to measure under natural conditions. Findings from magnification studies might give cause and direction for reorienting the existing framework of program planning. Question-

naires must take into account the possibility that respondents are unwilling or unable to describe their feelings or the reasons for their behavior. Time, money, and effort should be devoted to the development of research methodology uniquely suited to the evaluation of education programs.

The principle of "market segmentation" takes into account the fact that pupils vary greatly, and it can be used to learn more about the kinds of pupils that respond to particular programs so that programs can be tailored to pupils rather than mass implemented. While research can indicate which programs are financially most efficient, the relative value of the various measurable gains are matters of policy.

Weiss, Carol H. *Evaluating Action Programs*, Boston: Allyn and Bacon, 1972, in press.

This is a reader consisting of papers drawn from a variety of program fields on principles and strategies of effective evaluation. The emphasis is on problems that confront the evaluator in action settings and ways that have been found useful to deal with them. Among the contributors are Campbell, Riecken, Rossi, Stake, Scriven, Suchman, Weiss and Rein, Glennan, Freeman and Sherwood. The author's introduction considers and contrasts the readings and develops a coherent framework.

Weiss, Carol H. "Evaluation of In-Service Training," in *Targets for In-Service Training,* pp. 47–54. Washington, D.C.: Joint Commission on Correctional Manpower and Training, October 1967.

This article discusses evaluation of the outcomes of training programs: the definition of the objectives, specification of the objectives in behavioral terms, and measurement of the extent to which the objectives are achieved. Evaluation research should be done only when an agency really wants to know how good a job the training is doing in order to improve it. It is essential that the purposes of the training be specifically identified and their possible effects be faced up to as soon as the idea of conducting an evaluation is considered.

Evaluation of training is a three-part analysis. (1) Do the trainees learn? Have they shown changes in knowledge, attitude, and predisposition to apply new knowledge? If not, the

training program should be changed. (2) Do the trainees put their training into practice? If not, the agency should be examined for barriers it may present to the utilization of the training. (3) Are the trainees who practice what they learn more successful? If not, the agency should reexamine what it is teaching.

The three sets of questions may be answered by indicators like trainees' opinions, changes in trainees' knowledge or attitudes, changes in job performance, changes in client outcomes. The evaluation must deal with the kinds of acts that the training is designed to produce in order to be useful.

Weiss, Carol H. *Evaluation Research*, New York: Prentiss-Hall, 1972, in press.

This book is a comprehensive introduction to methods in evaluation research. Written in jargon-free style, it considers the range of issues in the field: clarification of the purposes that evaluation is to serve, formulation of evaluation questions, measurements of outcome, specification of the content of the program being evaluated and its component parts, research design, data collection, relationship between evaluators and practitioners, useful procedures when the program shifts course during the study, and the use of the evaluation results for policy and program development.

The book describes many of the newer developments in evaluation, such as quasi-experimental design, cost-benefit analysis, social indicators, planning-programming-budgeting, and structures and systems for improved dissemination of evaluation results. It relates these techniques to traditional evaluation methods and considers their advantages and disadvantages and the appropriate function that each can fill.

Weiss, Carol H. "Planning an Action Project Evaluation," in *Learning in Action,* edited by June L. Shmelzer. Washington, D.C.: Government Printing Office, 1966, pp. 6–21.

This article discusses the requirements for planning a social scientific evaluation. They include a commitment to serious evaluation by the program; early understanding of the use to which evaluation results will be put in order to determine its proper focus; clarification of project goals; the definition of goals in behavioral

terms; the choice of issues for evaluation; methods of selection of target group members; clarity about the target population; use of control groups; the securing of baseline data prior to beginning of the program; specification of program inputs; unambiguous measures of outcome; appropriate scope of expectations; attention to unexpected outcomes; appropriate statistical methods; and follow-up beyond the end of the program period. Other factors influencing the evaluation are discussed including the Hawthorne effect on the target group, evaluator-practitioner relationships, costs of the evaluation, and the value of both inside and outside evaluations. Evaluation research has significant limitations and often has little to say. But given skill, time, and resources, evaluation research can address itself to the questions of how much effect, why, and what else should be done.

Weiss, Carol H. "The Politicization of Evaluation Research," *Journal of Social Issues,* vol. 26, No. 4 (1970), 57–68.

This paper describes the growing visibility of evaluation research and its entry into the political arena. The politicization of evaluation research makes the evaluator vulnerable to methodological criticism, particularly from partisans whom his results displease; this makes him wary of using designs less traditional and accepted than the experimental model, even when they are more appropriate for the purposes of the study. Other effects are closer supervision of evaluation research by government funding bodies, more friction with program personnel, problems in drawing conclusions and making recommendations for future action, especially when the data provide little clear direction for change. A major problem is that evaluations tend to come up with largely negative findings.

Evaluators can play down the political implications of their work by (1) doing comparative evaluations that assess the relative effectiveness of different program strategies, rather than go-no/go evaluations; (2) avoiding premature evaluations of programs in flux, or (3) using system-model rather than only goal-model approaches.

But in a basic sense, the negative results that emerge from evaluation studies over a whole

range of programs are important data. They probably indicate serious shortcomings in the way social action programs are conceived, planned, and run. Basic social science has not provided many answers; program developers do not use much of the available information; administrative of programs is often deficient; programs are uncoordinated and provide fragmented services in trying to cope with broadrange problems. Bold new approaches are called for.

Weiss, Carol H. "Utilization of Evaluation: Toward Comparative Study," in *The Use of Social Research in Federal Domestic Programs,* vol. 3, pp. 426–432. U. S. Congress, House Committee on Government Operations, Research and Technical Programs Subcommittee, 90th Congress, 1st session, April 1967. Washington, D.C.: Government Printing Office, 1967.

Obstacles to objective evaluation of action programs have been vividly described, but some able evaluations survive the rigors of the environment. Technically competent and relevant to the issues, they still have had indifferent success in achieving either the discard or modification of ineffective programs or the institutionalization and spread of effective ones. This paper considers possible reasons for failure to apply results of sound evaluation and suggests systematic study of conditions associated with high and low utilization.

Lack of use is often blamed on the vested interests of policy makers and program operators. Characteristics of the evaluation itself are also significant. Among them are: (1) the extent to which evaluation addresses the underlying theoretical premises of the program, rather than only its operation in a particular place and time, (2) the extent to which it tracks the intervening processes through which effects are supposed to be obtained, (3) whether it is go/no-go evaluation, or analyzes the effectiveness of components of the program or alternative approaches, and gives leads to the kinds of change required, (4) whether results are positive or negative, and how drastic are the implied changes in philosophy, cost, staffing, structure, (5) the extent to which policy and program personnel participate in the evaluation process, and (6) the audience to which the evaluation is di-

rected—practitioners, administrators, higher policy makers, professional schools, or clients, each of whom has different motivation and capacity to implement results.

Knowledge of the effects of such factors on utilization can lead to improvement in evaluation practice and may help to overcome disenchantment with evaluation as a means of applying social science to the solution of social problems.

Weiss, Robert S. and Martin Rein, "The Evaluation of Broad Aim Programs Difficulties in Experimental Design and an Alternative," *Administrative Science Quarterly*, vol. 15, No. 1 (1970), pp. 97–109.

It is assumed by researchers that the ideal study design for evaluation of social programs is the controlled experiment. A case study is presented of an experimental evaluation of a social action program to make community institutions more useful and responsive to the community. The evaluators encountered technical difficulties and intra-organizational friction, which are virtually inherent in an experimental design for the appraisal of the effects of a broadaim, largely unstandardized and inadequately controllable action programs. Satisfactory criteria are difficult to find; comparison cases do not and cannot constitute real controls; treatments are not standardized; conflicts arise between evaluators and program directors. A plea is made for alternative models which are more qualitative, involve more informal interaction of evaluator with the target group and program directors, use observation and documents, and take account of political processes. Possible approaches for such process-oriented evaluations are: qualitative research, historical research, and/or case studies or comparative research.

Wholey, Joseph S., et al. *Federal Evaluation Policy*. Washington, D.C.: The Urban Institute, 1970.

This report describes federal practice in evaluation and makes recommendations for improvements. A particularly interesting section discusses the level of responsibility for evaluation.

Responsibility should be lodged at the decision-making level in order to avoid a situation where program managers must judge their own programs. Final evaluations (go/no-go and allocation of resources) should be made at the White House/Bureau of the Budget level. Responsibility for conducting national impact evaluations should rest with the agency head. For evaluation of projects within programs, and evaluation of the relative effectiveness of different program techniques, it is appropriate for operating bureaus and program managers to be in charge. When Federal programs are state and locally administered, the Federal agencies should retain at least some control over evaluations which are relevant to allocative decisions, and they should provide standards for local evaluations.

Wilder, David, "Problems of Evaluation Research," in *An Overview of Adult Education Research*, edited by Edmund deS. Brunner, David E. Wilder, Corinne Kirchner, John S. Newberry, Jr., pp. 243–273. Chicago: Adult Education Association of the U.S.A., 1959.

This article relates some of the major theoretical and methodological problems of evaluation research to the field of adult education, and makes recommendations about the design and conceptualization of studies.

The author notes several general problems which arise during the course of evaluating adult education programs. First, it is often difficult to define the goals of adult education, and they may be very different for the administrator of the program and for the participant. Since the goals are often diffuse and general (e.g. "cultural development"), operationalizing them may be a difficult task, particularly if a number of different courses are to be evaluated in the same study. Too narrow a conception of goals, such as measuring only increased aptitude in each course area, may ignore more important general goals, such as increased self-respect or participation in the wider community. Because there is usually such a wide variety of goals for the program, some of them are likely to be incompatible. Thus, it is important not only to define goals carefully, but to indicate the frame of reference which is used. Another major problem is that because adults are exposed to such a wide variety of stimuli, and because adult education programs are usually of fairly marginal importance in their lives, it is difficult to ascertain whether changes are due to the program or to other factors.

Studies in several areas related to adult education are reviewed, such as attendance, participation and dropouts, methods evaluations, and community education programs. Problems and general principles emerging from each case are discussed. Dropout studies, for example, have illuminated the problem of different frames of reference by showing that many people dropout not because they are dissatisfied with the course, but because they have learned what they wanted to. Studies of public information campaigns have shown clearly that merely exposing people to information will not make them learn. Methodological problems, such as obtaining a control group for widescale educational campaigns, are also discussed.

Recommendations for future evaluations include: (1) careful conceptualization of the evaluation model and the dependent and independent variables; (2) rigorous specification of techniques and instruments, and development of new measuring devices if older ones are unsuitable; (3) care in choosing the type of design which will produce reliable results. Many evaluations have been content to use the after-only design, which is usually inappropriate to the educational situation. It is also noted that far too many evaluations assume equivalence between the experimental and the control group without testing for it. (4) The author also warns against "overinterpreting" the data, and makes some comments about the development of interpretive hypotheses.

Wilkins, Leslie T. *Evaluation of Penal Measures*. New York: Random House, 1969.

This book examines the problems of evaluation with particular reference to penal measures. It discusses the varying goals of penal institutions, specifically treatment and punishment, and the concepts of efficacy and morality. It reviews the literature on recidivism and prediction methods. Current claims to knowledge about effective penal measures are listed. Techniques of analysis and the logic of inference necessary to evaluation in this field are examined in detail. The discussion is generalizable to evaluation in other areas of social action as well. Particular attention is paid to defining goals, allowing for error, operational sugges-

tions, choosing the right data, and experimental design in the context of an evaluation study that assesses the degree of achievement of program objectives.

Wilkins, Leslie T. "Evaluation of Training Programs" in *Social Deviance*, London: Tavistock Publications, 1964, pp. 288–293.

An alternative to the use of standardized tests in the evaluation of training programs is proposed. Standardized tests are insensitive to changes in course content and are unrelated to student characteristics. Since the purpose of evaluating training is to provide information for program improvement, this insensitivity is critical. One method proposed involves obtaining information (through open-ended essay questions) about trainee's attitudes, knowledge, reasoning processes, etc. Tests developed from this material are administered at the beginning and end of the program. The collected material is also used as a focus for the training content, within the program goals. Thus, the framework of the course is geared to the student's needs, and pre-test, post-test comparisons will provide information about the effectiveness of the program in areas of the trainee's specific weaknesses.

Williams, Walter, "Developing an Evaluation Strategy for a Social Action Agency," *Journal of Human Resources*, vol. 4, No. 4 (1969), pp. 451–465.

This article is concerned with the methodological and institutional problems faced by a social action agency in trying to make evaluation and important input to its decision-making processes. A developer of an evaluation strategy faces a discouraging set of problems: inferior methodological tools, severe field problems in implementing evaluations and even more severe problems in implementing new program ideas derived from evaluations, problems of integrating outcome evaluation results into the agency decision-making process, a basic weakness of pilot programs in producing good outcome data, etc. An agency's bureaucratic and administrative structure often resists evaluation, blocking both the development and the use of data. The best hope for the future lies in more competent staff, adequate time and funding, and increased

cooperation between evaluators and decision-makers.

Williams, Walter and Evans, John W. "The Politics of Evaluation: The Case of Head Start," *The Annals of the American Academy of Political and Social Science,* vol. 385 (September 1969), pp. 118–132.

The author, in this analysis of the controversy which arose over the Westinghouse-Ohio University evaluation of Head Start, emphasizes the conflict between the desire to produce large-scale, effective programs and the need to plan and analyze. The Head Start evaluation was conducted within a complex political framework. Head Start had expanded enormously from what was originally conceived to be a limited experimental program: it was publicly very popular and was thought to be effective. Because information for overall assessment was desired quickly, it was decided to do an *ex post facto* evaluation rather than a more rigorous but time-consuming longitudinal study. Program directors in the Head Start program opposed it on the grounds that the design, the weakness of available test instruments, and the failure to include any goals (such as health and community involvement) besides improved school performance would produce misleading results. The need for evaluation overrode these objections.

The evaluation of the effectiveness of Head Start in terms of intellectual and socio-personal development showed that summer programs appear to be ineffective, and full-year programs showed only marginally effectiveness. After the Head Start graduates had spent several years in school, they were considerably below national achievement norms.

The authors discuss criticisms of the study and conclude that although some are valid, the study is a relatively good one; most of the criticisms have been forthcoming because of the program's great popularity. Wider longitudinal studies would be better and should be done but this study provides at least some guidelines for decision-making at the present time.

A number of inferences are drawn from the larger issues inherent in this controversy: (1) there should be more skepticism about the capacity to mount large scale effective programs, particularly in those areas where the

main goals is to improve an individual's capacity to earn or learn; (2) a high priority should be put on efforts to restructure ongoing programs, or create new ones on a small scale, and to test these ideas before they are widely put into effect; (3) evaluation programs need to be expanded to improve the base of decision-making information; (4) difficult problems associated with the potential misuse of evaluation results remain, but it is more dangerous to launch large-scale programs which are untested.

Wilner, D. M., R. P. Walkley, T. C. Pinkerton and M. Tayback, *The Housing Environment and Family Life.* Baltimore: Johns Hopkins University Press, 1962.

This study was designed to test the widely held assumption that improved housing leads to improvement in physical, mental, and social health. The public housing project into which the test subjects were relocated provided the advantages of fewer persons per room, adequate heat, hot water and bathroom facilities, screens on windows and doors, adequate refrigeration and garbage disposal facilities. The study population came from the Baltimore Housing Authority's file of applicants for a new public housing project. All the families were Negro residents of Baltimore, mostly of lower socioeconomic class, and living in deteriorated slums. The final test and control samples were 300 pairs matched on 13 demographic items.

It was hypothesized that incidents of illness would be reduced. Test families were expected to be more likely to express satisfaction with housing, engage in more intra-family activities, have more favorable contacts with neighbors, have enhanced self-concept and higher aspirations, show increased participation in community affairs, and manifest improved psychological states. Superior school performance was anticipated for children through improvements in home environment and reduced absence due to illness. In eleven waves of interviewing over three years, mortality was found significantly greater in the control group, 10 vs. 2. Morbidity data in general confirmed the hypotheses for persons under 35, but not for persons over 35. Accidents were reduced by one-third in the housing project. However, data on personal and family relations were inconclusive. Three out of ten psychological scales

showed significant test-control differential. Hypotheses of school performance were not borne out on standardized test scores, but project children were more likely to be promoted on schedule and had great regularity of attendance.

Wittrock, M. C. "The Evaluation of Instruction," *Evaluation Comment,* vol. 1, No. 4 (1969), 1–7. (To be published in M. C. Wittrock and D. E. Wiley, *The Evaluation of Instruction: Issues and Problems.* In Press).

Evaluation studies must enable us to relate instruction to individuals and to learning. In order to evaluate instruction, it is necessary to measure at least three parts of instruction: (1) the environments of learning, (2) the intellectual and social processes of learners, and (3) the learning. Only then can the relationships among these three parts of instruction be quantitatively estimated. The reason why most previous evaluation studies have been of so little use to decision makers is that they have actually evaluated only one segment of the instructional program, and thus have not been able to point out the cause-and-effect relationships.

One form of evaluation has concentrated on assessing the quality of the educational environment in terms of books in the library, school budget per pupil, etc. Alone, this is clearly an inadequate measure of the quality of instruction, but characteristics of instructional environments should be made explicit in evaluation studies.

The evaluation of learners usually involves describing individual differences among students in abilities, achievement, and preferences. Such information provides the evaluator with a baseline from which to gauge the extent of impact, but does not enable him to make rigorous inferences about what the students have learned nor about the role of environments and intellectual processes in producing the learning.

Evaluation of learning has usually been done with standardized achievement tests. Such tests are often inadequate for evaluation because they focus on individual differences, and thus miss areas where all students have learned somethings at a high or low level. In using tests, it must be remembered that if students get a valid and reliable item wrong on a pretest, it is the instruction, and not the test, which must be changed. Behavioral objectives must be clear-

ly defined before test evaluations begin. The results should be judged individually for each student, against absolute and not relative standards.

Finally, the evaluation of instruction attempts to relate information obtained by the three sub-evaluations described above. This will usually involve multivariate statistical techniques and the development of models which can help the researcher to find the cause-and-effect relationships. Several books which have developed approaches to casual explanations from non-experimental data are recommended. The possibility of using path coefficients is particularly promising.

Worthen, Blaine R. "Toward a Taxonomy of Evaluation Designs," *Educational Technology,* vol. 8, No. 15 (1968), 3–9.

Stufflebeam's structure for generating evaluation designs (*Educational Technology,* vol. 8, No. 14), does little to simply the decisions which the investigator must make in each specific case. It would be optimal to have a taxonomy of evaluation designs relevant to different types of studies (context evaluation, input evaluation, etc.), but the complexity of the problem makes such a development unlikely at the present time.

Interim approaches at codification, however, are possible and two are suggested. First, lists might be made of alternatives for each decision situation. This would be useful but unwieldy. Second, a taxonomy of information needs might be developed. This could serve as a spur to the identification and development of a range of evaluation designs relevant to each need. When such multiple taxonomies of designs are developed, generalizations may emerge across classes of needs.

Wright, Charles R. and Herbert H. Hyman, "The Evaluators" in *Sociologists at Work: Essays on the Craft of Social Research,* edited by Phillip E. Hammond, pp. 121–141. New York: Basic Books, 1964.

This paper follows the processes involved in the conceptualization and conduct of an evaluation of "The Encampment for Citizenship." The design of the study focused on multi-wave questionnaires to participants and alumni. The authors stress the attempts they made to gain the trust of the administrators of the program

and to involve them in the planning and conceptualizing of the evaluation.

Control groups were set up by testing campers who were unable to attend, and by using different forms of the questionnaire for different groups and campers. Detailed lists of time schedules and chronologies are given. (For greater methodological detail, see Hyman, Wright, and Hopkins, *Applications of Methods for Evaluation: Four Studies of the Encampment for Citizenship.*)

Wrightstone, J. Wayne, Samuel D. McClelland, Judith I. Krugman, Herbert Hoffman, Norman Tieman, and Linda Young, *Assessment of the Demonstration Guidance Project,* Board of Education of the City of New York, Division of Research and Evaluation, n.d.

The goals of the Demonstration Guidance Project were to identify and stimulate able students from socially and economically deprived urban neighborhoods in New York City. The program provided increased guidance and remedial help, as well as broadened cultural exposure during both the junior high school and high school years. A main concern was to encourage the children selected to think about attending college. The project went on for six years.

The design of the evaluation was limited by the fact that it was not ethically possible to deny half the children in the demonstration schools the treatment program for such a long period of time. The control group selected consisted of the three classes that had graduated previous to the beginning of the project. Although there was clearly some limitation on comparability because of historical events, it is argued that the control group nevertheless was more comparable than a concurrent group selected from a different neighborhood in the city.

Scholastic ability was measured with traditional tests, and a positive gain relative to national norms was found in all areas, although reading scores were the most improved. The scholastic averages of the project students were considered very favorable when compared with similar students who had graduated from the school before the project was initiated.

The educational aspirations of the students did not appear to change greatly over the project period. These findings are somewhat dubious because the first measurement of aspirations was taken after the start of the program. There was, however, a 9.% increase in the number planning to go to college. When compared with a national sample of high ability students, the demonstration project students scored somewhat higher both on their aspirations and expectations of attending college. This was especially true of girls. They were, furthermore, more inclined to think that they would end up in a professional career than the national sample. Overall, the vocational and educational aspirations of the project students were very high.

Comparisons are made with national surveys, non-project high school students, and the control group mentioned previously. All of these show that the project students were superior. This does not, however, constitute proof that their superiority was due to the effects of the program. Nor does it indicate which aspects of the complex program were effective.

Another defect in the evaluation is the use of means to develop predictors of academic achievement, with no consideration given to the variance of scores within the compared groups. Moreover, it is clear from the results that standard tests alone are far less effective as predictors than are counselor and teacher ratings. The authors recommend the use of multiple measures for the purposes of prediction.

Evaluations of teacher, parent, and student feelings about the program showed that the overwhelming majority of all groups were enthusiastic, and felt that the program had had a demonstrable effect.

END