

THE USE OF GUIDELINES IN EVALUATING THE EFFECTIVENESS OF CLINICAL
LABORATORY IMPROVEMENT PROGRAMS, A META ANALYSIS

by

Deborah Joan del Junco

A thesis submitted to the faculty of The
University of Utah in partial fulfillment of the requirements
for the degree of

Master of Science

Department of Medical Technology

The University of Utah

August 1980

Copyright © Deborah J. del Junco 1980

All Rights Reserved

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

SUPERVISORY COMMITTEE APPROVAL

of a thesis submitted by

Deborah Joan del Junco

I have read this thesis and have found it to be of satisfactory quality for a master's degree.

Date

Co

Chairman, Supervisory Committee

I have read this thesis and have found it to be of satisfactory quality for a master's degree.

Member, Supervisory Committee

I have read this thesis and have found it to be of satisfactory quality for a master's degree.

Sarah A. Wise

Member, Supervisory Committee

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

FINAL READING APPROVAL

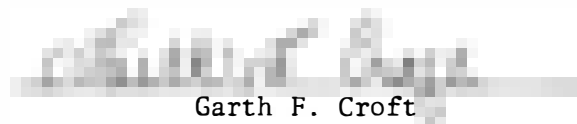
To the Graduate Council of The University of Utah:

I have read the thesis of Deborah Joan del Junco in its final form and have found that (1) its format, citations, and bibliographic style are consistent and acceptable; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the Supervisory Committee and is ready for submission to the Graduate School.



Member, Supervisory Committee

Approved for the Major Department



Garth F. Croft
Chairman/Dean

Approved for the Graduate Council



James L. Clayton
Dean of The Graduate School

ABSTRACT

Formal clinical laboratory improvement programs have been seeded by incendiary reports of fraud and error in diagnostic laboratories and the accompanying fear of public outrage. These programs are perpetuated by an intuitive notion that they foster quality health care, and that without them, conditions would be intolerable. The real efficacy of clinical laboratory improvement programs is debatable. When improvements in performance have been documented, the evaluation designs have not supported ironclad causal inferences.

This retrospective research examined the technical adequacy of 23 proposed and two completed evaluations of federally funded clinical laboratory improvement programs. The review process used throughout this research is referred to as meta analysis, which is a categorical term that means evaluation of evaluations or evaluation audit. Proficiency testing, technical consultation, and training were the three general approaches to laboratory improvement. A checklist of 31 evaluation guidelines was developed for the purposes of the review and for future use by program directors and funding agencies.

The data indicate that federally funded laboratory improvement programs continue to use technically weak evaluations. There were no significant differences in overall technical adequacy

between the three types of programs. However, there were significant differences between types of program proposals on 13 of the 31 individual checklist items. Eight of the 13 items were directly related to differences in requirements among the funding agent's three requests for proposals. The results suggest that the funding agent is in the best position to raise the technical quality of laboratory improvement program evaluation so that valid inferences as to program impact can be made and potentially worthwhile programs can be perfected.

CONTENTS

	Page
ABSTRACT.	iv
LIST OF TABLES.	viii
ACKNOWLEDGMENTS	ix
Chapter	
1. INTRODUCTION.	1
Laboratory Improvement Programs.	2
The Problem-Effectiveness.	3
Theoretical Framework.	5
Research Goals and Procedures.	5
2. LITERATURE REVIEW	8
Evaluation Theory, Design and Measurement.	8
Meta Evaluation.	20
Evaluation Applied to Continuing Health Professional Education	21
Evaluation of Clinical Laboratories and Laboratory Improvement Programs.	31
3. A FORMATIVE META ANALYSIS OF EVALUATIONS PROPOSED BY LABORATORY IMPROVEMENT PROGRAMS UNDER CONTRACT WITH THE FEDERAL GOVERNMENT.	41
Sample Selection and Characteristics	42
Checklist Guidelines	46
Analysis and Results	100
Conclusions.	112
4. A SUMMATIVE META ANALYSIS OF EVALUATIONS COMPLETED BY FEDERALLY FUNDED LABORATORY IMPROVEMENT PROGRAMS.	120
Selection of Two Example Programs.	123
Evaluation Example I	124
Evaluation Example II.	135
Summary.	146

Chapter	Page
5. SUMMARY AND DISCUSSION.	148
Review of Goals, Purposes and Procedures	148
Technical Quality in Laboratory Improvement	
Program Evaluation	150
A Need for Valid Data in Laboratory Improvement	
Policy--The Role of the Funding Agency	151
Implications	153
APPENDIX.	161
REFERENCES.	163
VITA.	173

LIST OF TABLES

Table	Page
1. Components of Program Evaluation	50
2. Ratings of the Evaluation Characteristics of Twenty-Three Laboratory Improvement Program Proposals.	102
3. Major Checklist Item Deficiencies.	104
4. CDC Proposal Specifications.	107
5. Associations between Type of Proposal and Checklist Compliance as a Function of RFP Requirements	110
6. Association between RFP Requirements and Differences in Compliance to Checklist Items	111
7. Changes in Laboratory Performance in Chemistry Post Technical Consultation.	129
8. Changes in Laboratory Performance in Immunology Post Technical Consultation.	130
9. Changes in Proficiency Test Error Rates Over One Year	133

ACKNOWLEDGMENTS

The initial stimulus for this work came from my dear friend, colleague, mentor and committee member, Dr. Nancy Coldeway, who is the Coordinator for Instructional Design at the InterWest Regional Medical Education Center in Salt Lake City, Utah. Several years ago Nancy aroused my curiosity in the field of evaluation, and ever since then she has added fuel to the fire. I am grateful for the inspiration which flows from even her most casual conversation, for her encouragement, and for her discerning judgment always delivered with the utmost diplomacy. To the other members of my committee, Dr. Sarah Wise and Ms. Sue Cockayne, I am indebted for their understanding, insight, and most of all their earnest endeavors to amplify the quality and usefulness of this thesis. Special thanks go to Dr. David Bradford from the Utah Poison Control Center for the major influence that his expertise in research design and data analysis had on this thesis.

I sincerely appreciate the benevolence of the Center for Disease Control Bureau of Laboratories and Negotiated Contracts Branch which made this research possible. The opportunity to work with CDC professionals like Dr. John Krickel, Dr. G. Richie Elwell, and Dr. Pegi Brooks (under earlier CDC contracts) was an enriching experience. Their genuine concern for worthwhile laboratory improvement programs instilled in myself and many others a resolve to find

the most effective approach; this thesis is really the culmination of a three-year collaborative search spawned and nurtured by CDC.

Finally, I wish to acknowledge my husband Gerry, whose love and approbation during this project has meant more to me than anything else.

Chapter 1

INTRODUCTION

That deficiencies in the quality of laboratory service exist is an undisputed truth. Even under the best of circumstances, random laboratory error occurs about one to three percent of the time (Sealfon, 1976). Considering that in 1978 the nation spent 12 to 14 billion dollars on laboratory tests, with annual increases running 15 percent (Relman, 1979), it is unlikely that the public will be sympathetic to even legitimate sources of laboratory error. To make matters worse, circumstantial evidence of laboratory deficiencies pervades the professional literature. Some of it surfaces in the mass media, in a dramatized version, and inspires congressional discourse (Finkel and Miller, 1973; Fouty, Haggen and Sattler, 1974; Javits, 1979; Kaufmann, 1973; Kauffman, 1979; McCormick, Ingelfinger, Isakson and Goldman, 1978; Sherman, 1979; Schaeffer, Widelock, Blatt and Wilson, 1967; Schoen, Thomas and Lange, 1971; U.S. Congress, Senate Committee, 1977; Wallace, CBS "60 Minutes" Report, "Do Medical Laboratories Need Tighter Control," 1979). The end result of this cycle is legislative action and a market for clinical laboratory improvement programs (Clinical Laboratory Improvement Act, 1967; Notice of Proposed Rulemaking, 1979; Peddecord, 1978).

There are anecdotal reports of misutilization of laboratory testing and serious iatrogenic injuries as the result of flagrantly poor laboratory work. These reports, along with apparent widespread fraud and corruption, have prompted a number of government, consumer, and professional organizations to press for higher quality and control of this subindustry within the health care field. (Peddecord, 1978, p. 1)

Laboratory Improvement Programs

The response to the demands for laboratory accountability has been variegated. To date, 13 states have instituted laboratory licensure laws, but only one of these requires individual licensure for laboratory workers (Forney, Blumberg, Brooke, Eavenson, Gilbert, and Kauffman, 1979; White, 1979). Mandatory personnel standards that affect various types of laboratories have been imposed by 20 states (Kull, 1980).

Certification is required of laboratories serving patients covered by Medicare and Medicaid (U.S. Department of Health Education and Welfare [DHEW, 1978]). Certification and accreditation (mandatory or voluntary) are the same in that they both entail laboratory personnel requirements, quality control standards, participation in proficiency testing programs, and periodic inspections (College of American Pathologists, 1974; Joint Commission on the Accreditation of Hospitals, 1976). Proficiency testing in this context refers to the distribution of simulated patient samples to laboratories "to determine their ability to achieve the correct analysis" (Forney et al., 1978, p. 128).

Training, continuing education and technical consultation programs, whether private or government sponsored, also seek to meet

the challenge to upgrade laboratories. For all practical purposes, training and continuing education are identical; individual laboratorians are presented learning materials, e.g., lecture, literature, visuals, and simulations, which are expected to be transferred to improved job performance. Continuing education may include activities that are personally interesting besides those that are functionally necessary. Training usually implies only the latter type of activity. In a 1967 conference on Manpower for the Medical Laboratory, Calvin Plimpton offered this reaction to the semantic bifurcation:

A spirit of curiosity is an attitude typical of good physicians, good nurses, good technologists, and I am sure it is this attitude which will be most responsible for progress in the future. This attitude can be stifled when people receive only training. . . . If you only train somebody, he will be left out in the cold if you introduce new procedures and new techniques. If however, he has been educated to think, he has acquired certain patterns of thought, certain ways of establishing qualitative judgments and therefore has the background to live with change and himself encourage improvement. ("Manpower for the Medical Laboratory," 1967)

Technical consultation is educational, but it is more of an ad hoc laboratory improvement effort than training, proficiency testing or accreditation. It usually involves an onsite visit to a laboratory where conditions underlying a problem can be observed and a specific course of action suggested (Schaeffer, Widelock, May, Blatt and Wilson, 1970).

The Problem-Effectiveness

The public can rest assured that a major campaign has been set in motion to combat impropriety and incompetence in clinical

laboratories; and the cost is on the same grand scale. Scores of potentially ameliorative laboratory improvement programs are in a very uneasy position; for all the money spent, they do not know for sure whether they have been effective (Peddecord, 1978). Carlson (1977) contends that evaluation of the effectiveness of laboratory improvement programs consists primarily of the personal bias of the author, modified only by some "ground rules of discussion" (p. 203). If this is the case, laboratory improvement programs boasting of success can be very beguiling indeed.

The situation is regrettably analogous to the state of affairs in recent evaluation research on the effectiveness of the Professional Standards Review Organizations (PSRO's) which were funded by the U.S. DHEW in 1972 and charged with promoting effective and economical delivery of health care services. At the request of the Subcommittee on Oversight of the House Committee on Ways and Means, the Congressional Budget Office analyzed PSRO programs for their effectiveness. Their 1979 report states:

Most extant evaluation studies are too flawed to be reliable, and furthermore, they yield inconsistent evidence. . . . Unless changes are made soon in both implementation and evaluation, future evaluations of the program will continue to be unreliable--often to such a degree as to be useless in formulating policy. ("Effect of PSRO's," 1979, pp. ix-x)

There may be some consolation in knowing the evaluation outlook is equally dismal in other professional circles outside health care. Taylor-Fitz-Gibbon and Lyons-Morris (1978a) discussed several disappointing studies of educational evaluation, one of which reviewed "2,000 projects that had received recognition as

successful . . . not one with an evaluation that provided acceptable evidence regarding project success or failure" (p. 12).

The rationale for this research stems from a clear need for valid evidence of the effectiveness of clinical laboratory improvement programs. Such evidence will provide a basis for sensible decision making. If there are no dependable data, the situation invites emotional arguments to dictate policy.

Theoretical Framework

Although evaluation theorists differ widely in their preference for evaluation designs, most agree that the primary purpose of evaluation is to guide rational decision making and facilitate value judgments. This differs from research whose purpose is to contribute to a body of knowledge (Alkin, Daillak, and White, 1979, p. 13; Cooley and Lohnes, 1976, pp. 2-3; Gephart, n.d.). Yet in order to fulfill its purpose, evaluation must borrow from research methodology and operate within the context of the scientific method, as a systematic process of disciplined inquiry (Anderson and Ball, 1979, p. 125; Rossi et al., 1979, Chaps. 5 and 6; Worthen and Sanders, 1973, pp. 10-14).

Research Goals and Procedures

The purposes of this thesis are to develop evaluation guidelines and to articulate the operational framework for valid evaluative inquiry into clinical laboratory improvement programs. It is hoped that the advantages of attending to the guidelines prospectively, before program implementation, will become apparent.

The ultimate goals of this research are to promote rational decision making and to channel creative energy into worthwhile laboratory improvement programs of mutual benefit to both health care consumer and provider. The following procedures will be carried out to achieve the purpose and goals:

1. The literature on general evaluation theory, evaluation in Continuing Health Professional Education and Clinical Laboratory Improvement Programs will be reviewed to introduce basic evaluation concepts and to trace their application to quality assurance and continued competence in the health care delivery system and more specifically, laboratory service.

2. Evaluation guidelines will be synthesized from several authoritative sources. The guidelines will be recategorized, assembled into a checklist, thoroughly described, and adapted to meet the needs of clinical laboratory improvement program evaluation.

3. The checklist will be field tested on 23 proposed clinical laboratory improvement programs that have been federally funded. The proposals will be rated on the checklist items to assess the technical quality of their evaluation plans. The results will provide direction for improving future clinical laboratory improvement programs.

4. Two completed laboratory improvement program evaluations will be reviewed to epitomize the subtleties of validity and invalidity in evaluation and measurement. This analysis will pick up where the checklist leaves off, to track the full gamut of laboratory improvement program evaluation from plans to practice to

conclusions. The pitfalls in the evaluation process will be uncovered so that future programs can avoid them.

5. The implications of the guidelines and suggestions for further research will be discussed to expedite the diffusion of technically sound, valid evaluation not only into the health fields, but throughout education and human services as well.

Chapter 2

LITERATURE REVIEW

Several substantive areas of the literature were reviewed. They are discussed under the following headings (1) evaluation theory, design and measurement, (2) meta evaluation, (3) evaluation applied to continuing health professional education (CHPE), and (4) evaluation of clinical laboratories and laboratory improvement programs.

This chapter begins by discussing very general evaluation concepts and proceeds to the more technical issues of design, measurement and analysis of behavioral attributes. Meta evaluation is described in detail since it is the general category of evaluation activities most relevant to this research. Finally, the chapter reviews evaluation design and measurement principles within the limited contexts of continuing health professional education and laboratory improvement programs.

Evaluation Theory, Design and Measurement

The theory underlying evaluation is discussed in this first section. Also included is a brief history of evaluation research traced from the 1960's, when the prolific works of a few authors elevated the status of evaluation to a growth industry; to the present, where evaluation can be seen as a complex mosaic resulting

from an effusion of sophisticated models (Rossi, Freeman, and Wright, 1979, pp. 24 and 27). Laboratory improvement programs have only to follow the precedent already set by educational and social science evaluation theorists.

Evaluation Theory

In the 1960's, individual initiative along with impetus from the federal government spawned two divergent schools of evaluation theory intended to aid decision making. L. J. Cronbach advocated measurement of post-treatment performance of a single well described group, while J. C. Stanley campaigned for rigorous experimental design using randomly selected and assigned treatment and control groups for comparative analysis (Hamilton, 1977). Michael Scriven somewhat tempered the disagreement by suggesting that multiple treatment groups be exposed to varying levels of educational intervention; thus no one would be denied treatment and valid comparisons would be possible. He also advised using multiple criterion measures to prevent overlooking possible program effects (Hamilton, 1977; Worthen and Sanders, 1973).

Both Cronbach and Stanley's concepts of evaluation are based on the specification of a goal or hypothesis. In deference to this goal-based orientation, Scriven pursued his own line of reasoning resulting in the development of a much broader concept, goal-free evaluation (Borg and Gall, 1979, pp. 603-605). The central figure in the goal-free approach is an unbiased evaluator who seeks to measure program effects in terms of what is good for the

nation as opposed to goals preset by a program director (Hamilton, 1977). Evaluation encompasses a group of activities that are carried out under either a goal-based frame of reference, where the evaluator compares what actually happens to a preconceived notion about what was expected to happen; or a goal-free frame of reference where the evaluator just observes what happens, without any expectations. Scriven is also the originator of the terms formative (developmental) and summative (final outcome) evaluation (Worthen and Sanders, 1973, pp. 60-104). Messick (1967) elaborated on Scriven's philosophy by urging evaluators to consider the environmental as well as achievement variables or matrix of traits that moderate an individual's learning. Despite all this eclecticism, the original schools of evaluation thought remained polarized on the issue of design.

Evaluation Design

There are basically two general classes of evaluation designs: experimental, where variables other than those to be manipulated are controlled by random selection and random assignment of subjects to a treatment group; and nonexperimental, where extraneous variables are not necessarily controlled because some nonrandom selection or assignment process is used. The term quasi-experimental has been applied to those evaluation designs somewhere in between (Campbell and Stanley, 1963). Uncontrolled extraneous variables have been grouped into several categories under the general rubric threats to internal validity (Borg and Gall, 1979,

p. 522; Campbell and Stanley, 1963). Unless the threats to internal validity are eliminated or controlled, the conclusions about the effectiveness of the treatment (or intervention) in bringing about change will be extremely vulnerable to disconfirmation. Only true experimental designs can support cause and effect conclusions beyond a reasonable doubt (Borg and Gall, 1979, p. 519).

There are also threats to external validity which plague the generalizability of study findings beyond the participants included. In this case, experimental designs do not have any particular advantage over quasi experimental designs and do not necessarily outshine nonexperimental designs with regard to external validity (Cook and Campbell, 1976, p. 299). A comprehensive and lucid treatment of the experimental design topic can be found in either Campbell and Stanley (1963) or Cook and Campbell (1976). Cook and Campbell's work considerably elaborates on Campbell and Stanley's discussion of internal and external validity. Internal and external validity will be further explicated in Chapter 3 to relate the particular threats to validity to commonly used laboratory improvement program evaluation designs.

Glass and Worthen (Worthen and Sanders, 1973, pp. 221-224) presented an engaging defense of experimental design in response to Guba and Stufflebeam's apparent condemnation. The repartee began with Guba and Stufflebeam's claim that the use of rigorous experimental design precludes the flexibility that is essential to program improvement during implementation. Glass and Worthen then replied that as long as an educational treatment creates an identifiable

context, an experimental design will allow flexibility and adaptation of the program to the exigencies of the moment.

An additional problem with experimental design, according to Guba and Stufflebeam, is its inability to control all the extraneous variables that come into play in educational evaluation. Randomization can never assure equal groups they asserted; though randomization may work in the experimental laboratory, it is not appropriate to the real world. Glass and Worthen conceded that the use of experimental design cannot absolutely equate two groups, but probabilistic comparisons are possible. Pertinent to this issue are some observations about experimental designs in field research made by Cook and Campbell in 1976. They seemed to justify some of Guba and Stufflebeam's contentions by describing how there are problems maintaining a control group in a field setting, e.g., treatment eventually diffuses into the control group and the control group's performance is adversely affected by their resentment from being left out of the treatment. Spurious results can thus plague even a rigorously controlled design (pp. 228-229). Nevertheless, Cook and Campbell seem to support Glass and Worthen in their exhortation of experimental design. Without it, they caution that inferential statistics cannot be correctly applied and internal validity suffers greatly.

Another approach will be mentioned here, not because of its contribution to evaluation design, but because of its detraction from it. It is a nonexperimental model often couched in the language of true experimental design. Campbell and Stanley (1963, pp.

64-71) categorized this type of evaluation as ex-post-facto correlational. It has also been referred to as post hoc or tacked on evaluation to cast doubt on any pretense of causation it may display (Dixon, 1978). The evaluator using this model draws a conclusion about the effects of a program based on the performance of the treated group compared to the performance of what she/he would like the audience to believe is a control group. In fact, the two groups should not be compared because their constituents have not been randomly assigned nor in any way matched for relevant characteristics. The control group may include those who opted not to participate or those who were unable to participate. The reasons they did not participate can be expected to influence their performance just as much as the treatment affects the participants' performance. For example, it has been shown that people who volunteer for behavioral research are usually better educated, more motivated, more altruistic, and more sociable than nonvolunteers (Rosenthal and Rosnow, 1975). If the program under study is a training course for laboratory workers intended to increase concern for quality control and skills in troubleshooting, without ever receiving the training, those who volunteer would probably outperform those who decline. To compare post-course performance of the volunteer group to the group that refused and infer that training improves performance would be sheer delusion. The antecedent conditions for good performance, in this instance, are motivation, altruism and some background education--the very traits the control group lacks.

The debate about experimental design versus more practical

approaches to evaluation has launched a whole cadre of evaluation theorists and practitioners in pursuit of the ideal model. For the purposes here, model and design are synonymous. To date, at least 44 models of the evaluation process have emerged (Carroll, 1980) largely from the efforts of 43 influential theorists and a core group of six individuals.

Hamilton (1977) has recently traced the backswing of the pendulum--away from experimental design which educators so vigorously espoused in the late 1960's and early 1970's. He delivered a persuasive case for an evaluation approach he referred to as pluralism. This model is mostly concerned with actual program activity and tends to downplay goals and hypotheses about expected or desired activity. Thus pluralists would be expected to operate under the goal-free evaluation theory. Pluralist evaluators use more of a magnifying glass approach, and employ a vast armamentarium of evaluation tools to detect program effects. Participants and providers are closely scrutinized as they engage in program activities.

The pluralism model includes the intense evaluation pursuits recently described by Smith (1978), i.e., educational ethnography, participant observation and case study. For pluralists, the individual not the institution, is the experimental unit. Pluralism, says Hamilton (1977) is best characterized by its expression of doubt and reflection in contrast to the plunge-ahead certainty and action of other models.

Whether the evaluator chooses an experimental, quasi

experimental or one of the nonexperimental evaluation designs, the measurements must be accurate and the statistical analyses of the data must be appropriate.

Measurement

A crucial decision every evaluator faces is the selection of relevant behavioral indicators and suitable instruments to measure them. Selection of evaluation criteria, particularly performance rating scales, should include consensus techniques and task analysis (Pierleoni, 1978; Wigton, 1980). The sophisticated field of psychometrics has evolved in response to the need to develop mental test instruments and other techniques to identify and quantify behavioral attributes (Nunnally, 1978, Chap. 1). The psychometric issues pertinent to this research include validity and reliability of measurements, and methods to explore relationships among measured variables.

Test scores mean very little unless they can be shown to correlate highly and reproducibly with the underlying behavior of interest. For example, if one wishes to measure intelligence, the test items must accurately discriminate between highly intelligent and not so intelligent individuals. If they do, the test items can be assumed to be valid. If individuals repeat the test, and achieve almost the same scores, the test would appear to be reliable. The concepts of validity and reliability of measurements are of central importance. These are discussed in Gronlund (1976, pp. 79-104) and Nunnally (1978, pp. 86-113) and will be elaborated on briefly here

because of their importance to the checklist in Chapter 3.

There are three major categories of validity: criterion-related, content, and construct validity. According to Gronlund, there are two elements of criterion-related validity: predictive and concurrent validity. However, Nunnally (1978) makes the point that their logic and procedures are exactly the same. Criterion-related validity refers to the degree to which a criterion measure accurately predicts performance in some other closely related dimension. An appropriate example would be the degree to which laboratory compliance to inspection standards predicts performance on proficiency tests, or an even more important consideration is whether proficiency test performance accurately predicts typical performance on routine patient specimens.

Content validity has to do with how representative a measure is of a particular domain of behavior or knowledge. A written test has content validity if the test items are matched to the course objectives and the content taught. To have adequate content validity, there must be enough measurement items to cover the domain of the content and the items must precisely relate to the topic.

Construct validity is extremely important to the measurement of abstract variables or constructs such as intelligence and problem-solving ability (Nunnally, 1978, pp. 94-109). According to Nunnally, construct validity involves specifying a domain of observables (p. 98), determining whether the observables tend to measure the same thing or different things, and determining whether the results of the measurement of the observables support the theory

underlying the dimension of behavior. An issue requiring special attention to construct validity is whether licensure of laboratories and of laboratory personnel assures quality in health care.

There is one major factor which can affect all three types of validity. Smith and Glass (1977) referred to it as reactivity or bias of measurement. Reactivity relates to the likelihood that a respondent will fake an answer to a test or survey item, or that an observer will unconsciously misinterpret a response. Often the participants in a study can guess the response desired, particularly on attitude inventories, and answer in the manner they think they should, rather than candidly. This is especially true if they wish to win approval from the evaluator or observer. Similarly, the observer who is rating responses or behaviors will often perceive things according to his or her personal bias. Some measures are inherently more reactive than others in their tendency to elicit distortion or falsehood. Smith and Glass (1977) rated the reactivity of certain measures used to evaluate psychotherapy outcomes. Their rating scale had five levels with physiological measures and grade point average as the low (most favorable) end of the scale and therapist's non-blind ratings at the high (undesirable) end of the scale. Posavac (1980) adapted the criteria and rating system to studies of patient education programs. His adaptation comes closer to having implications for continuing health professional education and laboratory improvement programs. The categories are listed as follows in order of increasing bias or reactivity:

1. Physiological measures and objective tests.

2. Objective variables that can be (but are unlikely to be) greatly affected by awareness that the participant is being evaluated.

3. Standardized tests of subjective states like emotional mood and personality traits.

4. Variables likely to be influenced by the participant's desire to appear favorable to the evaluator.

5. The evaluator's non-blind ratings of knowledge or compliance.

Reliability of measurement deals with repeatability or "random influence which tends to make measurements different from occasion to occasion" and is also affected by measurement error (Nunnally, 1978, p. 225). Reliability can be estimated by examining internal consistency via coefficient alpha formulas (Nunnally, 1978, p. 230), the Kuder Richardson 21 formula, analysis of variance and split half techniques (Gronlund, 1976, pp. 108-112). Inter-rater reliability is of particular importance in laboratory evaluation. Inter-rater reliability "is easily determined by correlating scores obtained from different scorers on the same and alternative forms of the measure" (Nunnally, 1976, p. 232).

Psychometrics is concerned not only with scores on individual variables, but also the way single variables relate to one another (Nunnally, 1978). The technique of multiple regression, for example, can specify a set of best predictors from among many performance measures for a particular dependent variable (e.g., patient health status, accurate laboratory test results, etc.).

Carrying multiple correlation techniques one step further leads to multivariate analyses. One particularly useful multivariate technique, factor analysis, can elucidate constructs and reduce large numbers of related variables to more manageable factors (Nunnally, 1978, Chaps. 10 and 11). Multiple regression can be performed on factor scores (from factor analyses) and is currently a very popular practice in educational research with important implications for studies of laboratory improvement as well (Kukuk and Baty, 1979).

This section has reviewed the validity and reliability principles of measurement to provide a foundation for their proper application in laboratory improvement programs. A brief discussion of methods to explore relationships among different variables was included to introduce the multivariate concepts which will be discussed in the subsequent sections of this chapter and again in Chapter 4.

The foregoing discussion of evaluation theory, design and measurement covered the major evaluation issues in a generic sense to provide the necessary basis for applying the key concepts specifically to continuing health professional education and laboratory evaluation. This literature review also introduces terms that will be used in the evaluation guidelines in Chapter 3.

One other general term will be discussed here before proceeding with the sections on continuing health professional education and laboratory improvement. This entire thesis is built upon the concept of meta evaluation. Meta evaluation can be seen as a

milestone in the logical progression of evaluation thinking.

Meta Evaluation

Meta evaluation is, in essence, an evaluation of evaluation(s). Scriven originally coined the term meta evaluation in 1969, but others have contributed to the development of the concept, and to further refinement and advancement of its applications (Scriven, 1976, pp. 133-134; Stufflebeam, 1978), particularly Stufflebeam. Meta evaluation, in Stufflebeam's interpretation, is a concept that relates to the assessment of the merit of a particular evaluation.

Stufflebeam (1978) indicated that meta evaluation is a valuable consumer protective device. It is a systematic way to assess the extent to which an evaluation is technically adequate, useful in guiding decisions, ethical and practical in its use of resources. Meta evaluation can uncover the strengths and weaknesses of a single study or an entire group of evaluations. Studies can be compared to one another (Hamilton, Baker, and Mitchell, 1979) or aggregated to determine the overall effectiveness of a general class of educational intervention (Posavac, 1980; Smith and Glass, 1977). Glass (1977) cautioned that in the latter sense, meta evaluations should serve descriptive rather than inferential purposes due to vast complications surrounding valid statistical analysis of aggregations of studies. Since this study evaluates the strengths and weaknesses of laboratory improvement program evaluations, it can be considered a meta evaluation.

Evaluation Applied to Continuing Health
Professional Education

Evaluation of laboratory improvement programs can gain considerable insight from a review of past and present evaluation in continuing medical education, continuing nursing education and continuing education for allied health professionals. The following paragraphs refer to continuing education in these areas in a generic sense employing the term continuing health professional education (hereinafter referred to as CHPE) to encompass all health fields.

Annual expenditures for continuing health professional education (CHPE) exceed the billion dollar mark. Eventually, the health care consumer will bear the burden (Lloyd and Abrahamson, 1979). Meanwhile, its effect on the quality of health care remains an enigma, perhaps due to the fact that impact evaluation of CHPE is one of the most underfunded areas of research (Lloyd and Abrahamson, 1979; "U nursing study," 1979).

Current evaluations of CHPE programs rarely seek patient health-status data or evidence of participant behavior change. They are more likely to examine such easily measured variables as postcourse satisfaction or knowledge gain (Connelly, T., 1979). Program evaluators are liable to make an unwarranted inferential leap if they conclude that high satisfaction ratings lead to improved patient outcomes (Berg, 1979; Dixon, 1978; Newstrom, 1978). Empirically based estimates of the predictive validity of satisfaction ratings and cognitive tests are needed to legitimate such

intuitive impressions.

The purpose of this section is to examine the state-of-the-art of continuing education evaluation and to consider the attendant methodological problems that bear important implications for laboratory improvement program evaluation.

CHPE Evaluation Designs

Unlike educational evaluation in public schools, CHPE evaluation is totally dependent on volunteer participation. This poses a serious threat to external validity (Rosenthal and Rosnow, 1975) in terms of generalizing results to the entire target population of health professionals. Restriction to volunteers also exacerbates the usual difficulty associated with the randomization process that is essential to internal validity (Campbell and Stanley, 1963, p. 5). This could explain why Lloyd and Abrahamson (1979) were able to find only two continuing medical education programs out of 47 reviewed, that used random assignment in their evaluation designs. This paucity of rigorous design is consistent across other health professions as well (Dixon, 1978). Campbell (1967, p. 283) offered a possible solution to the volunteer problem. Twice as many volunteers as can be accommodated can be recruited to attend a CHPE program. Randomization will then decide who will be enrolled into the program and who will not. This may jeopardize external validity if the members of the control group react with indignation to the restricted enrollment, and perform worse than they normally would on the criterion measure. However, if they can

be offered the program at a later date after the evaluation, and they are informed of this when they originally volunteer, there should be no problem.

In 1976, Inui, Yourtee and Williamson were able to randomize treatment to intact groups in a matched-control quasi experimental design. They offered convincing evidence that their continuing medical education program on hypertension succeeded in improving physician practice and patient health. An ingenious evaluation of a continuing education program for pharmacists employed a trained observer to pose as a patient in a contrived situation. The investigators showed positive long-term effects of their program compared to a valid control group (Dixon, 1979). However, the evaluators repeated their observations 12 months later only to find that the participants (from the treatment group) had regressed. This kind of valid data, though discouraging on the surface, is extremely important in that it demonstrates a need for reinforcement to maintain improved behaviors.

Without an equivalent group for valid comparison, the conclusions drawn about program effectiveness are vulnerable to many alternative explanations. From a political or ethical standpoint, it may prove impossible to leave the selection of who gets a potentially ameliorative program up to chance (Cook and Campbell, 1976, pp. 300-301). Again, Campbell (1967) offered a possible strategy referred to as staged introduction (pp. 279-281). A program can be distributed to comparable intact-groups at successive intervals. Those who haven't yet received the program provide an

adequate control group for comparison.

When a no-treatment control group is to be used, internal and external validity will be increased if the control group is given a placebo or Hawthorne control (Borg and Gall, 1979). This reduces the probability of spurious significant differences between treatment and control groups due to one group perceiving itself as special while the other feels neglected; or the probability that real significant differences will be obscured by the control group's competitive desire to perform equally as well as the treated group--also known as the John Henry Effect (Borg and Gall, 1979, pp. 162-164).

Other CHPE evaluation approaches--mostly of the ex-post-facto type--dominate the literature. Many CHPE programs are more goal-oriented than research oriented. Evaluators of CHPE often confine their inquiry to whether or not a program goal has been achieved for one single group at one particular point in time. This involves a simple case study of the participant group, a pre and posttest of the group, or a variety of other nonexperimental methods. A recent study by Walsh (personal communication P. Walsh, InterWest Regional Medical Education Center, Salt Lake City, Utah, June 3, 1980) compared goal-based to goal-free evaluation in examining the effects of a continuing education program for a heterogeneous group of health professionals. Walsh found that the goal-based evaluator actually met fewer preset evaluation criteria than the goal-free evaluator. Goal-free evaluation offers an attractive alternative to rigorous experimental design without sacrificing

objectivity. It does not eliminate as many threats to internal validity, however.

It is impossible to unilaterally extol all evaluations using experimental designs and condemn all ex-post-facto models. There are pitfalls surrounding the latter, however, that necessitate careful consideration. Since others have so aptly covered the subject (Campbell and Stanley, 1963; Cook and Campbell, 1976), one cogent example will serve the purpose here.

Page et al. (1979) reported a study of a continuing education program for physicians in which the control group and treatment group were self-selected. The investigators conducted an analysis of variance on group scores from several postcourse performance measures. It appeared that the program improved the performance of the treatment group ($p < .001$) while the performance of the control group declined. Results of a pretest showed that the mean of the control group was significantly higher than the mean of the treatment group. As illustrated, the posttest group means were the reverse of the pretest means:

	Pretest	Posttest	n
Participant Group	2.64	3.67	116
Nonparticipant Group	3.62	2.33	115

When the treatment and control group are nonequivalent, this kind of regression toward the mean often occurs (Borg and Gall, p. 523 and p. 591). Page et al. also discovered that the apparent

improvement in postcourse performance for the treated group could have been explained by a gradual national trend upward. These are but a few of the rival explanations that can challenge the results from an ex-post-facto design. Considering that most of the studies that reach the published literature are success stories (Posavac, 1980), this honest report of specious findings that disappeared under scrutiny encourages a skeptical perspective.

The state-of-the-art of CHPE evaluation probably does not lag far behind educational evaluation in general. CHPE and laboratory improvement program evaluators can learn much from reviews of evaluations in the public school systems (Hamilton, et al., 1979). The most common and persistent deficiencies there include lack of appropriate instruments to measure goal achievement, lack of testing for statistical significance, flawed evaluation designs, faulty data collection, and biased data management.

This section discussed how evaluation designs are being applied in studies of the impact of CHPE. There have been some attempts at rigorous evaluation design, however, the majority of continuing education programs reported in the literature rely on nonexperimental designs.

CHPE Measurement

Several measurement techniques have been used in assessing CHPE programs. This section will discuss four types of measurement in CHPE: participant reaction, learning and attitude change, typical behavior, and patient health status (Dixon, 1978; Newstrom,

1978). Possible ways to improve the construct validity of some of these measures will be suggested. In addition, several factors which can interfere with the full impact of CHPE programs on performance and patient health, will be described.

Unobtrusive, nonartificial measurement of typical on-the-job behavior and patient outcomes supplies the information most valuable for deciding whether a continuing education or training program has been successful, provided the evaluator has some valid basis for comparison (Borg and Gall, 1979, pp. 159-162; Webb, Campbell and Schwartz, 1966). Unfortunately, the evaluation criteria most frequently measured are participant reactions of satisfaction (Connelly, 1979; Dixon, 1978; Lloyd and Abrahamson, 1979; Newstrom, 1978). Though this information is useful to formative evaluation, studies show it does not appear to correlate with the effectiveness of the course in achieving its ultimate goals (McGuire, Hurley, Babbott, and Butterworth, 1964; Williamson, Alexander, and Miller, 1967).

Studies also show that performance on cognitive tests does not correlate with typical job performance (Loughmiller, Ellison, Taylor, and Price, 1970; Taylor, Price, Richards, and Jacobsen, 1964; 1965). The test-item format (true/false, multiple choice essay, etc.) has been shown to be one moderating variable (Benson and Crocker, 1979; Harasym, Baker, and Mitchell, 1979; Newble, Baxter, and Elmslie, 1979). There are probably many others such as level of motivation, extrinsic incentives, and attitudes. If competency based education and criterion referenced testing become more common

in the training of health professionals, the correlations between cognitive tests and measures of typical performance or patient health status may improve. "A criterion referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of a specified domain of instructionally relevant tasks" (Gronlund, 1976, p. 19).

There is also a discrepancy between attitude inventories and measures of typical performance (Dixon, 1978). Reliance on any kind of self-report is often a risky practice unless the evaluator is strictly interested in formative evaluation information (Olson and Fruin, 1979).

Even when the optimally objective measures of patient outcomes have been used, the results have been inconclusive and somewhat disappointing. Lloyd and Abrahamson (1979) found that of the four studies (out of 47) that related CHPE attendance to patient health status, only two reported improvements. Given that generalization from measures of performance to real attributes is not always valid (Jaeger, 1978), the final question remains unanswered: To what extent do any of these measures indicate the actual quality of health care? A large part of the CHPE evaluation problem appears to be a lack of construct validity (Engel, 1978).

To increase construct validity and provide a sensitive tool for monitoring both job performance and patient health status, some physicians have advocated patient chart audit and the involvement of the Professional Standards Review Organizations (PSRO's) for evaluating the adequacy of their CHPE programs (Caplan, 1973; Jessee,

Munier, Fielding, and Goran, 1975; Reed, Lapenas, and Rogers, 1973). This is a fairly unobtrusive method in that it takes advantage of already existing institutional records. However, many questions are being raised about this peer review method, especially in terms of the predictive validity of the process oriented audit criteria that are used to make inferences about patient outcomes. Fifer (1979) refers to this doubt as a period of reflection and introspection following the realization that PSRO proponents underestimated the difficulty of the task (measurement of the quality of care). This hesitant stance indicates a penchant for the pluralism evaluation model (Hamilton, 1977).

There is hope that the psychometric technique of factor analysis will greatly improve CHPE evaluation by its capability to clarify constructs (Dielman, Hull, and Davis, 1980; Engel, 1978). Evaluators may discover that while they think they are measuring a single construct, e.g., competence on the job, they are actually measuring more than one. The act of aggregating more than one construct totally confounds any meaningful interpretation of the results (Nunnally, 1976, Chap. 10). In fact, Davidge, Davis and Hull (1980) found--through factor analysis--that the criteria their faculty were using to rate medical students' clinical competence actually represented two different factors: interpersonal skills and problem solving skills. The two would have to be kept separate in an attempt to measure improvement at some later point in time, e.g., after an educational intervention.

Factor analysis has also been combined with multiple

regression to determine which factors best predict desirable outcomes ("Final Report," 1979). Such a capacity could certainly shed light on the PSRO process versus outcome controversy. There are methodological cautions important to this application of factor analysis beyond the scope of this review. They are well described in Kukuk and Baty (1979). A more immediate concern is the possibility of observer bias if the factors consist of observers' ratings. Webb et al. (1966) described 21 systematic sources of bias that apply to observer ratings. Observers must be trained; inter-rater reliability and internal consistency must be determined.

It is clear that there are definite obstacles to valid measurement of the impact of CHPE. Of even greater concern are the intervening and confounding variables that greatly dilute any measurable impact. They include the geographical, administrative, structural, and systems aspects of a health professional's work setting (Brown, 1977, pp. 11-18; Dixon, 1978; Jessee et al., 1975). A CHPE program may recommend changes that the participants have no authority to implement. If CHPE programs would prospectively take these variables into account and assure that the right people get the right program, more conclusive and perhaps more positive evaluation results might evolve.

The motives for attending and the educational backgrounds of CHPE participants are often diverse. Those who are already familiar with a course topic may tune out. Many professionals attend a course to expand their knowledge, not to correct a deficiency (Jessee et al., 1975). When this is the case, the impact

of the course may be too elusive to measure. If a contributing factor in the change process is participant motivation, it should be measured along with the other alleged performance indicators.

This section has reviewed the four types of measurement used in evaluation of CHPE programs. They were: participant reactions, learning and attitude change, typical performance, and patient health status. The use of PSRO's and statistical techniques such as factor analysis and multiple regression, was discussed as a possible means to overcome the lack of construct validity inherent in the most common measures of CHPE effectiveness. Finally, the obstacles and facilitators of CHPE program impact were described to illustrate how CHPE programs can best accommodate them.

Despite the obstacles, limitations, and sources of invalidity, more rigorous evaluation methods are possible and necessary. Better evaluation will lead to better ways to upgrade the quality of health care. Cronbach pointed out an added bonus (Worthen and Sanders, 1973, p. 47): "Eventually better evaluation will train better teachers." One final caveat is in order here (Worthen and Sanders, 1973, p. 231) to put the issues in proper perspective:

A poorly executed, premature and inconclusive comparative summative evaluation will only drain precious resources which could be spent more wisely on formative evaluation. . . . [However] eventually, a hard-headed summative evaluation based largely upon a comparative experiment must be performed.

Evaluation of Clinical Laboratories and Laboratory Improvement Programs

This section examines the indicators which have been used to evaluate the quality of laboratory service. These same indicators

are being used to evaluate the effectiveness of laboratory improvement programs. Organized laboratory improvement is discussed as it is conceptualized by its founders--as a conduit leading to competent performance and assurance of accurate, reliable laboratory data (Forney and Brooke, 1967; Schaeffer et al., 1970).

Clinical Laboratory Evaluation

Efforts to measure the accuracy of laboratory testing began at least as early as 1946, with the work of Belk and Sunderman (1947). Their survey of the accuracy of chemistry testing in 49 institutions is considered a landmark in proficiency testing (Forney et al., 1978). Their use of pooled sera to simulate real patient specimens provided the prototype for modern day proficiency test programs.

There are three functions of proficiency testing according to Forney and his group (1978, p. 149):

1. To provide each participant with a critical evaluation of the performance of the person's own laboratory.
2. To provide information to the profession about many aspects of laboratory science.
3. To provide data for regulatory purposes when this is required.

Others cite education as a fourth possible function of proficiency testing (see Bibliography in Appendix: Connecticut State Department of Health; Iowa State Hygienic Laboratory; Commonwealth of Kentucky Laboratory Improvement Program; Massachusetts Health Research Institute).

Forney et al. (1978, p. 151) elaborated on the second function of proficiency testing listed above. They pointed out that proficiency test programs monitor the state of analytic performance of laboratories, determine the closeness of agreement of laboratories (inter-laboratory reliability), and provide the data necessary to compare current performance levels to the needs of health programs (p. 151). The advantage of proficiency testing as a performance measure is its relatively low cost; its primary disadvantage is that it is a measure of maximum capability rather than typical performance (Peddecord, 1978).

Some research has been done to assess the degree of difference between typical performance and maximum laboratory capability. Three studies using blind vs. identified proficiency test specimens have found considerable differences in performance. Participants performed better when they were aware they were being evaluated (La Motte, Guerrant, Lewis, and Hall, 1977; Black, Dorsey, and Whitby, 1976; McCormick, Ingelfinger, Isakson, and Goldman, 1978). One study found no significant difference in blind vs. nonblind proficiency test performance (Steele, Schauble, Bechtel, and Bearman, 1977). The individual studies differ in methodology, e.g., types of laboratories studied, types of laboratory tests reviewed, and methods of introducing the simulated specimens. This variability renders any generalization across the studies highly speculative.

The validity of proficiency testing (PT) as a measure of laboratory performance and quality of care in general, has been challenged; the medical significance of laboratory errors has yet to

be incorporated into proficiency test scoring criteria ("Laboratory Proficiency," 1976; Peddecord, 1978, p. 35). Proficiency test scores in one laboratory discipline lack predictive validity for another discipline (Peddecord, 1978, p. 44). Low PT scores may occur due to low prevalence of the diseases that the PT specimens represent (Peddecord, 1978, p. 87).

Reliability is also a problem in proficiency testing. Peddecord (1978) called attention to the considerable variability in PT scores that may be found from one survey to the next, and from year to year (p. 48). The degree of difficulty varies with the composition of the specimen in areas requiring qualitative judgments.

It would appear that the domain of laboratory proficiency is sampled neither uniformly nor sufficiently when proficiency tests are assembled. Laboratory proficiency is apparently not one single construct. Aggregating scores across disciplines may confound several different constructs and attenuate their differences. This in turn will obstruct the formulation of rational laboratory improvement policy. Programs may adopt a shotgun approach when a far less expensive focused approach is indicated.

On-site inspections have also been used in laboratory evaluation. Checklists have been developed by several different agencies, i.e., the College of American Pathologists (1974) Joint Commission on Accreditation of Hospitals (1976), Medicare (U.S. DHEW, 1978) Food and Drug Administration, the Center for Disease Control, the American Association of Blood Banks, and individual State Health Departments (Garcia, K. W., 1980). The checklists are

similar in that they all call for yes/no decisions on the part of the inspectors. They are not uniformly consistent in their explicit standards or implicit judgment criteria (Forney, et al., 1978; Garcia, K. W., 1980).

The thrust of the inspections is in the direction of the laboratory's structure and processes underlying their test results or outcomes. Donabedian (1969, pp. 186-215) categorized the facilities, equipment, qualifications of personnel, organizational hierarchy and fiscal policy of the health care institution as variables of structure. He classified technical competence and protocols as process factors. The advantage of structure and process data over outcome data is "insight into the nature and location of deficiencies or strengths to which the outcomes might be attributed" (Donabedian, 1969, p. 188).

The assumption is often made that given the proper structure components along with a technically adequate process, high quality service will follow. However, the true relationship between structure, process, and outcome is complex, ambiguous and not yet understood (Donabedian, 1969, p. 207). Peddecord (1978) raised many questions concerning traditional measures of laboratory structure, process, and outcomes, when he found no significant relationships between PT scores and deficiencies noted during inspections in any clinical laboratory disciplines except Bacteriology.

In addition to an apparent lack of validity of onsite inspection checklists, interrater reliability has not been established (Peddecord, 1978, p. 84). There appears to be a lack of

standardized interpretation of regulations among Medicare Surveyors (U.S. Senate Hearings, 1977, pp. 590-618; Peddecord, 1978, p. 85; Sherman, 1979). The Center for Disease Control has recently held several conferences for Medicare Surveyors in order to reach consensus on interpretation and enforcement of regulations. The effect of the conferences has not yet been reported. Peddecord (1978) found that the ratings of individual CAP inspectors were at least fairly internally consistent within a discipline. However, there does not appear to be any predictive validity of checklist ratings across disciplines ("An Analysis of Idaho," 1979b, p. 14).

Other evaluation methods are possible. This review of the literature did not find any reports of their use in laboratory performance evaluation over the last twelve years. The possibilities include (but are not limited to): onsite observations of the laboratory in typical operation (non-artificial); achievement tests of laboratory personnel; self-reports of laboratorians' perceived training needs; attitude inventories of laboratory personnel; review of anecdotal records such as written complaints and commendations; critical incident reports; sociometric methods (Gronlund, 1976, Chaps. 16 and 17); reviews of quality control records; surveys of patient opinions; surveys of clinicians opinions; reviews of employee performance evaluations; interviews or surveys of opinions of other departments' employees; and patient chart reviews.

Peddecord (1978, p. 8) observed that the important communication, utilization and interpretation aspects of laboratory testing have been largely ignored by current laboratory evaluation

techniques. There is some isolated evidence of physician misutilization and misinterpretation of laboratory data (Casscells, Schoenberger, and Graboys, 1978; Hardison, 1979; Kassirer and Pauker, 1978; McGuckin, Adenbaum, and Corbin, 1979). Each of the laboratory evaluation measures presently in use extracts only a slice of reality from the much broader health care picture. They yield results inevitably distorted.

The ideal measure of the quality of laboratory service would be unobtrusive and would tap into the laboratory-clinician interface. The real patient population of the facility would be the stimulus, not artificial specimens concocted under arbitrary circumstances. The perfect measuring device would record precisely and calibrate accurately according to criteria of medical significance. Laboratory service would be evaluated as a composite of patient-clinician-laboratorian interactions.

Until the ideal is achievable, the validity and reliability problems of current laboratory evaluation techniques may be overcome by combining multiple data sources and measurement instruments (Peddecord, 1978; Donabedian, 1969). Evaluators should be prepared for some contradictions, however. For example, it has been shown that physicians often judge laboratory effectiveness by turnaround time, fees and test variety (Fouty, Haggan, and Sattler, 1974). These factors would not necessarily correlate with proficiency test success. Nevertheless, a multidimensional perspective would be more likely to perceive the complex reality of the total picture. When all of the variables affecting laboratory service are considered in

the laboratory's evaluation, the way will be paved for maximally effective laboratory improvement programs.

Laboratory Improvement

Proficiency testing, onsite inspections, consultation and continuing education have all been used to improve laboratory performance (Peddecord, 1978, pp. 11-12; Forney et al., 1978). Professional organizations, universities, state health departments, and the Center for Disease Control have been involved with these activities since at least as early as 1962 (Forney and Brooke, 1967). A few studies citing these methods have been reported in the literature (Sattler, 1970; Schaeffer et al., 1967; 1970; Fouty, Haggren, and Sattler, 1974). Quantitatively, their results either showed no change or were equivocal.

There are reports of laboratories' PT scores improving with prolonged enrollment in a PT program (Peddecord, 1978; Finkel and Miller, 1973). However, this does not necessarily indicate that proficiency test programs improve typical laboratory performance. A laboratory may persist in performing poorly on patient tests, while proficiency test scores show marked improvement. In order to demonstrate the effects of proficiency testing on laboratory performance, some external measure of performance other than PT is needed. Proficiency testing as a form of laboratory improvement or treatment is distinct from proficiency testing as a laboratory evaluation instrument. The same holds true for laboratory accreditation inspections.

The goal of any special program or treatment should be related to the needs. The real need in the case of laboratories is not better proficiency test scores or fewer inspection deficiencies; it is accurate, reliable, meaningful patient test results.

There is a tendency for programs to confuse problems with symptoms of the problem or solutions to the problem (Mager and Pipe, 1970, p. 2). Anderson and Ball (1978, p. 17) advised program originators to consider needs in performance-deficit terms not treatment deficit terms. Mager and Pipe cautioned that

. . . statements such as 'We've got a training problem' are pits into which one can pour great amounts of energy and money unproductively. Such statements talk about solutions not problems. Training is a solution . . . [that] implies transferring information to change someone's . . . ability to perform. But lack of information is often not the problem. (1970, p. 8)

Mager and Pipe discourage the use of the word deficiency in performance evaluation. It connotes unequivocally bad performance. They prefer the term discrepancy be used to avoid jumping to erroneous conclusions.

A difference between what someone is doing and what you would like them to be doing is not enough reason to take action. . . . We must be selective about which discrepancies to attack. The way to do that is to check the consequences of leaving the discrepancy alone. (Mager and Pipe, 1970, pp. 11-12)

Laboratory improvement programs and their evaluation plans should take into account those variables that mediate performance. For example, "Poor selection of techniques is an important factor in the low rate of acceptability of lab determinations" (Finkel and Miller, 1973). Another factor to consider is the effect of test volume or workload on laboratory proficiency. In health care, the

frequency of performance of a particular procedure has been shown to be significantly related to the outcome. Whether the task is surgery or laboratory testing, the more procedures performed, the better the results, up to a point (Luft, Bunker, Enthoven, 1979; Finkel and Miller, 1973). The more minute and peripheral details of a procedure or performance standard may require mnemonic devices, as McDonald (1974) observed in a study of physicians. And finally, the complex interactions and interdependencies of the laboratory, other members of the health care team, and the institution's administration, must be recognized as both facilitators and obstacles to the improvement and valid evaluation of clinical laboratory performance.

To summarize this chapter, the quality of extant evaluation design and measurement has undergone careful scrutiny in the general fields of education, social science, and health care. The literature is replete with the resulting critical reviews. Laboratory Improvement Programs, however, have not been examined for the adequacy of their evaluation designs. Such a review is conspicuous in its absence considering that laboratory improvement programs represent a substantial investment on the part of taxpayers, health care consumers, and health professionals. Without thorough analysis and constructive criticism, deceptive evaluations can continue with impunity, while meticulous evaluations decline due to lack of incentive.

Chapter 3

A FORMATIVE META ANALYSIS OF EVALUATION PROPOSED BY LABORATORY IMPROVEMENT PROGRAMS UNDER CONTRACT WITH THE FEDERAL GOVERNMENT

This entire Chapter and Chapter 4 can be considered meta evaluations in that they report on investigations of laboratory improvement program characteristics that indicate basic technical quality of the evaluations. The work of evaluation theorists has been reconstructed into the practical guidelines which constitute the checklist used for this chapter's meta analysis. The derivation of the guidelines is also the first step towards the development of a valid evaluation style for laboratory improvement programs. The guidelines can be used in their checklist format by future program planners. This follows Scriven's (1976, pp. 119-139) premise that the prospective use of checklists prevents evaluator's defensiveness and tendency to overlook relevant evaluation considerations, the two most common obstacles to first rate evaluations.

Twenty-three recently successful clinical laboratory improvement program proposals were examined according to a checklist of evaluation guidelines. The process and results of the analysis will be discussed in this chapter under the headings of (1) sample selection and characteristics (2) checklist guidelines (3) analysis and results (4) conclusions. To investigate how certain components of

the proposal writing process contribute to the technical quality of the evaluations planned, the following null hypotheses will be tested:

1. There is no difference in overall technical adequacy of the evaluation plan between the three different types of proposals, i.e., proficiency testing, technical consultation and training.

2. The technical quality of particular aspects of the evaluation plan (or ratings on individual checklist items) is no different from one type of program proposal to the next.

3. Any differences found among the three types of proposals, in terms of their compliance to particular checklist items, are not related to the variability of the stipulations in the funding agent's three requests for proposals (RFP's).

4. The amount of contract money awarded is not related to the overall technical quality of the programs' evaluation plans.

5. The amount of prior contract experience is not related to the technical quality of the programs' evaluation plans.

Sample Selection and Characteristics

The sample selected for the analytical review in this chapter consisted of twenty-three funded contract proposals. The proposals are listed in the Appendix. They represent the entire group of successful contractors who submitted program proposals in April 1979, in response to requests for proposals, hereinafter referred to as RFP's, circulated by the Center for Disease Control (CDC)

Negotiated Contracts Branch. Three different RFP's were distributed to all potential contractors including universities, state health departments and professional organizations. Each RFP called for a different type of laboratory improvement program. CDC designated the types as follows: proficiency testing, technical consultation and training. Although the general program categories were predetermined by CDC, individual proposal writers were encouraged to innovate within a given category. Prospective contractors were allowed to submit no more than one proposal for each category ("Center for Disease Control," 1979).

The sample was selected based on the following rationale:

1. Formal laboratory improvement programs under the auspices of the CDC have been in existence since 1962 (Forney and Brooke, 1967). Since then, CDC has had vast experience with laboratory evaluation, consultation and training programs. Thus they would be likely to award funds only to high quality, technically sound programs, which would incorporate the benefits of 17 years of evolution in the field of laboratory improvement into their proposals. Such programs could set the standard for the state-of-the-art of laboratory improvement program evaluation.

2. CDC contract funds have been awarded to 61 laboratory improvement programs since 1977, the first year that monies were made available (personal communication, Andrea Terrill, Center for Disease Control, Negotiated Contracts Branch, February 5, 1980). Eight of the 19 contractors funded in 1979 (the contract year of interest in this investigation) have each been awarded contracts

for the three consecutive years. Five contractors have each garnered funds for two years. Three contractors have each secured six contracts, one contractor has had five, and two contractors have each had four over the three years. Since greater than two-thirds of the 1979 contractors have had considerable experience with program evaluation, it was believed that the majority of 1979 contract proposals would bear a high degree of quality and sophistication in their evaluation approaches.

3. The contract proposals awarded in October, 1979 represent the most current federally funded laboratory improvement programs. The programs have only recently begun operation as of this writing. It was therefore deemed most useful to provide timely information on strengths and weaknesses of their proposed evaluation methods for the major potential audience of this research, i.e., the future contract contenders and the funding agent. Although no monies are available for 1980, there is a possibility that funding will resume in 1981 in the form of cooperative agreements (personal communication, R. Eric Greene, Assistant Director, Bureau of Laboratories, CDC, October 31, 1979). The results of this analytical review should be well-timed.

The nineteen contractors are geographically distributed as follows: one in the Northwestern United States, two in the Southern Rocky Mountains area, three in the Great Plains area, four in the Great Lakes Region, two in the Mississippi Valley, four in the Northeastern United States, and three in the Southeastern United States. The only region not represented is the area in the Far West

of the United States, although a contractor within this area was funded in a previous year. The nineteen contractors represent eighteen states.

Although it is not known how many or which region's contract proposals have been rejected, it is curious that the geographic distribution is so widely representative. Three speculations come to mind: representative geographic distribution is a higher priority for the awarding of funds than is the conceptual and technical quality of the program proposals; the number and quality of program proposals submitted from each region of the United States is consistent enough across all regions that even highly discriminating selection results in all regions being represented; or, the number of proposals submitted is not equivalent across regions, but by some rare chance event, the highest quality proposals happened to be submitted from almost every region of the United States. A fourth possibility is that the awarding of funds is based on a random selection process. This does not seem to be the case. CDC does have proposal reviewers who rate proposals on a weighted point system, and unsuccessful contenders receive notices of rejection citing lower ratings on technical quality as the reason (personal communication, Andrea Terrill, CDC, Negotiated Contracts Branch, February 5, 1980).

In summary, the sample consisted of the entire group of successful contract proposals funded by CDC in 1979. The proposals were selected for their timeliness, and the technical quality it was assumed they would demonstrate due to CDC's high standards and

the contractor's own experience with laboratory improvement programs.

Checklist Guidelines

The term checklist is used, rather than model, because it is more straightforward and connotes the practical nature of the approach. The term model connotes an underlying theoretical foundation and the use of the word is sometimes considered pretentious (Anderson and Ball, 1978; Shepard, 1977).

This section will discuss the sources from which the guidelines were derived and the organization of the items into general categories. The second half of this section will describe the checklist items in detail including the rating scales and examples of the two extremes on the scales.

Sources and Organization of Checklist Items

The checklist used consists of 31 items, adapted from a variety of sources. The work of the Evaluation Research Society (1980); Stufflebeam (1978); Hamilton, Baker, and Mitchell (1979); Smith and Glass (1977); Scriven (1974); Shepard (1977); and Posavac (1980) contributed substantially to the development of the checklist.

The Evaluation Research Society (ERS, 1980) has recently proposed 55 standards for use by program evaluators in many diverse fields including health, education, welfare, law enforcement, public safety, business, training, and licensing. The standards apply to six general categories of evaluation: front-end analysis or needs assessment, evaluability assessment (a systematic way to

determine whether one should evaluate a particular program), formative evaluation, impact or summative evaluation, program monitoring, and evaluation of evaluation or meta evaluation. Only front-end analysis, formative, impact, and program monitoring evaluation are incorporated into the checklist and will be defined and described in greater detail in the definition section of this chapter. The ERS (1980) standards are organized into six sections: formulation and negotiation, structure and design, data collection and preparation, data analysis and interpretation, communication and disclosure and utilization. The checklist used in this chapter's investigation only considers standards listed under formulation through data collection. Some of the other factors under the ERS categories of data analysis and interpretation will be considered in Chapter 4.

Stufflebeam (1978) chairs the Joint Committee on Standards for Educational Evaluation. Their primary mission is to develop comprehensive evaluation guidelines specific to education. Publication of the standards is time-lined for 1981. Although the official standards have yet to be released, Stufflebeam set forth his personal preferences in a recent article on meta evaluation (1978). As he holds the highest position on the Joint Committee, it is probably safe to assume that he will have considerable influence in the group's final decision. His priority list includes 34 standards; many were reflected in the subsequent publication of the ERS (1980). Stufflebeam grouped his standards into four general categories: technical adequacy, probity, utility, and practicality.

Technical adequacy refers to the truthfulness of conclusions drawn and the replicability of the study. Standards under the ERS categories of structure and design, data collection, and data analysis resemble Stufflebeam's technical adequacy standards. The standards under Stufflebeam's probity category focus on ethical issues as do several components of the ERS's formulation and data collection categories. Stufflebeam uses the term utility to group and characterize those standards that have to do with the same communication and disclosure issues discussed in the ERS document.

The ERS proposed standards and the standards Stufflebeam advocates provided the conceptual framework for the synthesis of the checklist in this research. Most of the individual checklist items have been distilled from a consolidation of these two works. Individual items have been rearranged into general categories under the technical aspects of evaluation. The general categories listed by Stufflebeam and the ERS were revised and adapted specifically to suit laboratory improvement program proposals. The contribution of other works (Hamilton et al., 1979; Posavac, 1980; Scriven, 1974; Shepard, 1977; Smith and Glass, 1977) to the checklist is synergistic with rather than independent from the above described standards. These individual works will be discussed later only as they relate to specific checklist items.

In summary, the checklist items can best be described as the result of an eclectic rather than original process. Of the many facets of program evaluation--conceptual, technical, ethical, economic and effectual--only the technical aspects were selected for

this retrospective research. For greater conceptual clarity, the general class of technical aspects has been subdivided into three categories: context, structure, and instrumentation. This subdivision and nomenclature is based on a logical relationship of the items within each category. The subcategories listed under the technical component are further divided into more specific points at issue. The overall program evaluation scheme is presented in Table 1.

Descriptions of Checklist Items

This subsection will proceed as follows: each general category, i.e., context, structure and instrumentation, will be defined first followed by a detailed description of the individual guidelines it subsumes. The description of each item will include an explanation of the judgment criterion for each point along the rating scale. The categories and checklist items will be discussed in the same order as they appear in Table 1. Only the technical components of the evaluation scheme (see Table 1, heading II) are considered in the checklist.

The individual items comprising the checklist will probably overlap in the actual implementation of evaluation. However, in the development and pilot testing of the checklist, it was discovered that many items had to be divided in two, sometimes three, to avoid confounding multiple characteristics within a single statement (Nunnally, 1978, p. 79). It was also preferable to have many items in order to maximize the precision of the ratings.

Table 1

Components of Program Evaluation

-
- I. Conceptual
Basic orientation of the program director and evaluator to evaluation theories and particular models

 - II. Technical--Organization of the evaluation reflecting sound evaluation principles and scientific methodology
 - A. Context
 - 1. Clear need
 - 2. Defined purpose
 - 3. Description of target setting
 - 4. Description of target population
 - 5. Program approach
 - 6. Plan for cooperation/public relations (PR)
 - 7. Assessment plan
--manifest needs
 - 8. Assessment plan
--perceived needs
 - 9. Replicable, exportable program
 - B. Structure
 - 1. Formative evaluation plan
 - 2. Trial run
 - 3. Program monitoring plan
 - 4. Impact evaluation plan
 - 5. Evaluation design
 - 6. Inferences intended
 - 7. Statistical tests
 - 8. Unit of analysis
 - 9. Method of selection/assignment
 - 10. Internal and external validity
 - 11. Unbiased evaluator
 - 12. Evaluation meets audience objectives
 - 13. Evaluation meets program objectives
 - C. Instrumentation
 - 1. Measurement Methods
 - 2. Identification of instruments
 - 3. Estimation of validity
 - 4. Estimation of reliability
 - 5. Judgment criteria/standards
 - 6. Reactivity of measurement
 - 7. Data management system

Table 1 (continued)

- 14. Provision to measure unintended outcomes
- 15. Plan to evaluate long term effects

III. Ethical

Demonstration of respect for health, welfare, and privacy of participants or recipients and awareness of limitations

IV. Economic

Cost benefit of the program and efficiency of operation

V. Effectual

Utilization of evaluation results; persuasiveness of evaluators and conduciveness of evaluation to decision making.

A. Context. The technical context of the evaluation relates to the underlying conditions and the general milieu in which the program and the evaluation will take place. The chain of events leading up to the installation of the program and the operational framework of the program are the relevant components of context.

1. Clear need. Description--need refers to the social importance and absence of substitutes that justify the program's existence (Scriven, 1974, pp. 7-33; Shepard, 1976, p. 10). It is essential to discuss the need in terms of its saliency in the health care delivery system. An attractive program may be desirable, but this does not qualify it as a necessary program. Poor performance may not be improved by even the most palatable training program if the poor performance is due to organizational obstacles or lack of motivation (Mager and Pipe, 1970). A proposal should describe the prevailing conditions in the clinical laboratories which clearly indicate a need for change, e.g., inaccuracy, erratic results, or some other problem that interferes with the delivery of reliable, medically useful laboratory data. The proposal must explain why its particular program will meet the need while other existing programs do not.

Rating scale--2 (optimum) = program is explicitly justified based on a socially important need. 1 (partially acceptable) = proposal implies justification or describes a problem of questionable importance. 0 (inadequate) = a clear need is neither explicit nor implicit in the proposal.

2. Defined purpose. Description--the overall purpose, goals, objectives and characteristics of the program that will lead to achievement of the goals are precisely stated (Evaluation Research Society, 1980, p. 11; Stufflebeam, 1978). The purpose should be directly related to the need or problem defined in the first item, clear need. A well articulated purpose will be an indispensable aid to the evaluation plan. Without a program purpose or with a feebly described one, the evaluation plan is doomed from the beginning. Rossi et al. (1979, pp. 64-65) encourage proposal writers to use action-oriented verbs, to limit objectives to one single aim per statement, and to indicate how results will be determined when setting goals for a new program.

Rating scale--2 (optimum) = purpose is clearly stated in proposal and related to need. 1 (partially acceptable) = purpose is weakly stated or only indirectly related to need. 0 (inadequate) = stated purpose is not measurable or not consistent with the underlying need for the program.

3. Description of target setting. Description--the environment in which the study and the evaluation will take place is described (Stufflebeam, 1978). The proposal should consider all of the environmental and ecological conditions that could moderate the implementation and effectiveness of the proposed treatment. This will be helpful in planning a long term evaluation and an evaluation strategy to uncover unintended outcomes. The proposal should specify at least the geographic

location; the communities served by the laboratory; the bed capacity of the facility if a hospital (or the approximate test volume if not); the accreditation of the facility; the types of tests performed; the departmentalization or specialization of the laboratories; the organizational hierarchy; the amount of automation available; and the laboratories' professional clientele, whether general physicians, nurse practitioners, etc.

Rating scale--2 (optimum) = proposal describes the institutional environment of the target population in detail. 1 (partially acceptable) = description of target setting is sketchy or incomplete. 0 (inadequate) = no description is provided.

4. Description of target population. Description--the individuals who will participate in the program are characterized. (Stufflebeam, 1978). Their educational background; experience; professional affiliations; age; sex; attitudes toward their work, co-workers and other health professionals; job descriptions; performance levels; and other variables are all important if they can in any way affect the implementation and effectiveness of the program. Individual differences are especially important to formative evaluation and to the evaluation of unintended effects. Some variables can be retained and used in data analyses to allow the program more insight into the effect of their program on one group of individuals vs. the effect on a different group. For example, the program evaluators may find that the performance of college graduates improved after they read the instructional material whereas,

the performance of high school graduates who had been trained on the job did not improve after they read the material, but did improve after they were shown videotaped demonstrations.

Rating scale--2 (optimum) = proposal describes personnel in sufficient detail so that the job qualifications and job responsibilities of the target population are apparent. Other variables should be included or at least examined before program implementation. 1 (partially acceptable) = proposal does not describe target population explicitly, but shows some intention to consider their characteristics. 0 (inadequate) = the attributes of the target population are not even alluded to in the proposal.

5. Program approach. Description--a program is considered innovative if its treatment or delivery of treatment is unconventional or unique. A program is considered standard if its treatments conform to the procedures that are in common use for its category (Hamilton et al., 1979). Innovative programs are often appealing to funding agencies, but they must be carefully and rigorously evaluated since little or nothing is known about their effects (Boruch, 1976). The funding agency will usually make it clear whether they are interested in the development of new approaches or whether their primary goal is to disseminate the standard approach.

Rating scale--in the case of federally funded laboratory improvement programs, CDC encouraged innovation and new approaches to laboratory improvement (personal communication, Richie Elwell,

CDC Bureau of Laboratories, Laboratory Management Consultation Office, May 10, 1978; John Krickel, Laboratory Training and Consultation Division, January 9, 1979). Therefore, 2 (most desirable) = proposal describes an innovative approach. 1 (moderately desirable) = elements of the proposed program are innovative, but the overall approach is standard. 0 (least desirable) = program will apply the conventional or standard approach to laboratory improvement without any novel modifications. The judgment criteria would be reversed if the funding agent preferred the standard approach over innovation. The proper rating scale for this item depends on thoughtful consideration of the intention behind the funding agent's impetus.

6. Plan for cooperation and public relations.

Description--cooperation from all program participants must be secured. A system for communicating with other influential organizations and individuals should be included (Stufflebeam, 1978). Some programs have to be marketed or energetically promoted (Scriven, 1974) before the target population will participate. In the case of laboratory improvement, many programs would not get off the ground unless they had been either mandated or made very attractive. Continuing education programs seem to have more appeal on the surface than performance monitoring programs like proficiency testing. A well conceived public relations plan will map out techniques to win support, generate interest, and establish rapport with participants and other groups which may be affected by the program, including

health care consumers.

The public relations methods available to a program include letters soliciting support, scheduled conferences with representatives of special interest groups, coverage in the mass media and advertising in professional publications, telephone recruiting of participants, membership or subscription offers, remuneration and special bonuses. Of course legislative mandates and regulatory precedent could be cited as the primary means to ensure cooperation, but public relations activities should be included in the program's human service posture.

Rating scale--2 (optimum) = proposal describes plan to engender support from all important groups and methods to be used to recruit participants. 1 (partially acceptable) = proposal describes only one or the other of the above. 0 (inadequate) = proposal neglects to describe both recruitment and public relations.

7. Assessment plan-manifest needs. Description--the real performance problems or deficiencies are identified. These assessment activities should take place before the program is installed (Evaluation Research Society, 1980), but not necessarily before the proposal is prepared. Mager and Pipe's (1970) compact monograph on analyzing performance problems would be an invaluable aid to program staff charged with this responsibility. Manifest needs are observable. Overt behaviors, products, and outcomes such as laboratory test accuracy and precision are examples of observable performance. This item is distinct

from clear need, in that manifest needs are much more specific. For example, the justification or clear need for a program might be that patient health is compromised by inaccurate laboratory data. The observed problems or manifest needs in the laboratories referred to in this situation might be consistently low proficiency test scores, poor attitudes toward quality control, or reports of errors on patient tests. The clear need for the program should be readily apparent, whereas manifest needs may require closer examination to be identified. Manifest needs are the underlying causes for an unacceptable level of health care, which is the clear need for the program's existence.

Proficiency test scores and Medicare surveyors are sources for uncovering specific problems, but perhaps even more meaningful indicators are complaints that may have come from patients or clinicians or other laboratory personnel, supervisors' judgments, clinician's reactions to questionnaires about laboratory service, and incident reports.

Rating scale--2 (optimum) = proposal outlines the plan to uncover performance deficiencies. 1 (partially acceptable) = proposal alludes to possible deficiencies, but no clear intention to measure their prevalence or severity is mentioned. 0 (inadequacy) = proposal makes no mention of performance deficiencies or plan to assess manifest needs.

8. Assessment plan--perceived needs. Description--the needs and priorities perceived by the program participants are

considered. This is similar to what Scriven dubbed market research (1974, pp. 12-13). The perceptions of the participants (or recipients) of the laboratory improvement program should be systematically collected. There are dual benefits to this kind of needs assessment. One is that the important insights, attitudes and preferences of the people who will be most affected by the program will have a better chance of being incorporated or at least represented in the planning stages of the program. The second profitable outcome of assessing perceived needs is that participants will realize that their ideas are respected and solicited. Getting participants involved this way may increase their commitment and reduce the imminence of attrition. M. L. Brooks (n. d., a, b) from CDC has published two manuals on analyzing both perceived and demonstrated needs that are particularly relevant to laboratory trainers. Even if the program provides proficiency testing rather than training, the importance of determining the market for the program should not be overlooked. Questionnaires and telephone surveys are the most common methods to solicit perceived needs from the target population. Interviews and round-table discussions are also possibilities. Hearsay and incidental comments do not qualify as indicators of perceived needs.

Rating scale--2 (optimum) = proposal outlines plan to collect target population's perceptions of their needs.

1 (partially acceptable) = proposal implies that some participants' opinions will be or have been considered, but no

systematic collection method is described. 0 (inadequate) = no perceived needs assessment plan is evident or implied.

9. Replicable, exportable program. Description--the program and its evaluation methods are described in sufficient detail to enable replication of the complete study in another similar environment (Hamilton, et al., 1979; Stufflebeam, 1978). In drafting a proposal, the prospective contractor should explain what the treatment will look like in operation; what instructional materials or simulated specimens will be used; the number and characteristics of people and laboratories which will participate; and what sort of people will disseminate the program in terms of qualifications and demeanor. The evaluation efforts should also be thoroughly outlined. The following descriptions should all appear in the proposal: what kind of design will be used and why, what measurements will be taken, how will instruments be developed or adapted, and what are the judgment criteria. Sample test items or simulations (as in PT specimens) should be included for optimum clarity. The proposed activities must be feasible for someone else to duplicate under the same conditions. Rossi and his group (1979, pp. 74-75) caution that a program must limit its objectives to variables that are manipulable. The program may be a valid means to correct real and serious problems, but the stated goals may be unrealistic.

Rating scale--2 (optimum) = proposal describes a realistic, feasible laboratory improvement treatment, evaluation design, and format for assessment instruments. 1 (partially

acceptable) = proposed program or evaluation plans appear impractical or overly optimistic, or the treatment and evaluation plans are somewhat incomplete. 0 (inadequate) = the treatment or evaluation is not described in enough detail so that someone else could develop and implement the plans.

To recapitulate the context category, its constituent items reflect the important features of the total program. The program context sets the stage for the structure items which make up the outward appearance or framework of the evaluation. The context is naturally established first and the structure follows.

B. Structure. Referring to Table 1, structure appears as the second category in the technical dimension of evaluation. The structure category encompasses the majority of specific technical guidelines. This is logical since "technical" in the evaluation sense refers to the organization of the program evaluation based on scientific principles, and organization implies structure. Just as the architect draws a blueprint for a building's construction, the evaluator diagrams the structure for implementation of the evaluation. The final appearance of the building can be unique and imaginative, but the plans conform to certain standards and axioms, otherwise the building will collapse. Evaluation, too, must follow certain laws so that the conclusions are supported. The analogy of the building also serves well to illustrate the relationship between evaluation structure and evaluation context. A building (structure) rests on its foundation (context); just as the structure of an evaluation is grounded on its context.

1. Formative evaluation plan. Description--this includes all of the "mechanisms whereby the product will be continually upgraded" (Shepard, 1977, p. 45). Formative evaluation is also referred to as process evaluation (Hayman and Napier, 1975) and developmental evaluation (Evaluation Research Society, 1980). The activities of formative evaluation involve systematic data collection. The results are used to reshape the program as needed in order to maximize the final or impact evaluation results. For example, a laboratory improvement program labelled training may find that students are bored with a formal lecture. This observation should instigate a search for alternative instructional strategies. Many such small but important changes can be made in a program without destroying the overall impact evaluation design. However, if the proposal stated that one of the program objectives was to determine the effectiveness of lectures as compared to group discussions, it would be irrational to change the instructional format midstream. Modifications must be made carefully, not capriciously. Evaluators should delineate contingencies and set a maximum threshold before modifying the program to any great extent, as is done in medical research to prevent control group individuals from being denied a new therapy that is shown to be highly effective shortly after the study's inception (Boruch, 1976). The bottom line here is that the program staff must be observant and responsive; and this requires forethought.

A proposal could describe any of the following formative

evaluation techniques: post-course student critiques to be used after every training session or technical consultation visit, periodic questionnaires to be distributed to program participants, observation of the program in action by an independent, impartial evaluator, telephone survey of participants, or program staff's observations (these should be combined with other data originating from participants). Written pre and posttests of participants' cognitive skills are an excellent source of formative evaluation information in training programs.

Rating scale--2 (optimum) = proposal specifies a formative evaluation plan including data collection methods, so that appropriate improvements in the program approach can be made while the program is in operation. 1 (partially acceptable) = proposal indicates some awareness of the possible need to improve the program along the way, but offers no clue as to how the need will be determined. 0 (unacceptable) = proposal does not suggest any possibility for flexibility in program implementation.

2. Trial run. Description--the Evaluation Research Society (1980) considers a trial run or field test an activity within the formative evaluation rubric. The trial run concept is kept separate in this checklist to assure that it will not be omitted from the evaluation plan since formative evaluation refers to many activities. Scriven (1974) stresses the necessity of field trials prior to mass dissemination. This allows the treatment (or service) to be refined and polished before it reaches the target population. Hayman and Napier (1975, p. 45)

distinguish between pilot study and field testing as two separate phases in the chronology of program development. A pilot study involves trying the program out under controlled conditions using participants who have characteristics similar to the real intended participants, but pilot-study subjects are not drawn from the real-participant population. A field trial is a trial run of a program in the real setting using a small group of typical participants from the target population. The word typical is emphasized to differentiate these participants from a subset of participants who may perform far better or far worse than the average. The ideal policy is to pilot test first, field test second, and lastly disseminate the program to all participants. However, programs limited by contract deadlines usually have time for only one type of trial run. A field test would probably provide the most valuable information.

Rating scale--2 (optimum) = proposal describes plan to field test or pilot test the program prior to mass distribution.
1 (partially acceptable) = proposal suggests that some component of the total program will be pilot tested, e.g., the evaluation instruments or one of several different training courses.
0 (inadequate) = proposal does not delineate any pilot test or field test plans.

3. Program monitoring plan. Description--the ERS points out that "this is the least acknowledged but probably most practiced category of evaluation" (p. 7). Program monitoring activities include counting the number of program participants,

the number of drop outs, or the number of complimentary and derogatory telephone calls received. A technical consultation program could plan for the consultant to count the number of questions asked by participants during the onsite visit in addition to the other counts of attendance, etc. A proficiency test program could plan to count the number of requests for replacement shipments which would provide an indication as to the quality of the service provided. Simple counts may seem rather mundane compared to more elegant data collections and analyses currently practiced, but unless some systematic documentation takes place, useful data of the highest reliability will go unnoticed. Webb, Campbell, Schwartz, and Sechrest (1966) have provided a classic work extolling the virtues and accessibility of unobtrusive, nonreactive measures; certainly head counts and enumerations of letters of praise or criticism qualify as unobtrusive measures of program effects. Many ingenious ways to monitor a program's implementation and development are possible if only some systematic method of recording is planned in advance (Rossi, Freeman, and Wright, 1979, pp. 38-40 and 122-157). Counts should focus on both positive and negative occurrences. Program monitoring must be sensitive to events that are detrimental as well as conducive to program goals. The information collected should be used to assure that the program is proceeding according to plan. The appearance of the program in operation should be periodically compared to the proposal and to established standards for the program (set by the funding agent, for example).

Rating scale--2 (optimum) = proposal describes plan to collect enumerative data including at least personnel attendance rates, number of laboratories participating, number of telephone calls and letters received concerning the program and number of participants who drop out of the program. 1 (partially acceptable) = proposal lists only one of the above data sources in their program monitoring plan. 0 (unacceptable) = no plan to monitor the program is described.

4. Impact evaluation plan. Description--impact evaluation is the best means to find out how well the overall program worked (ERS, 1980). Other names have been ascribed to the process such as Scriven's term, summative (Worthen and Sanders, 1973), and Hayman and Napier's reference to outcome evaluation (1975), but they all converge on the same basic concept: effectiveness. The information emanating from an impact evaluation can form the basis for the decision to award future funding or withdraw support (ERS, 1980; Hamilton, et al., 1979). Impact evaluation is therefore, an awesome and sometimes loathesome task for the hopeful staff of a fledgling program (Hayman and Napier, 1975, pp. 3-7; Lyons-Morris and Taylor-Fitz-Gibbon, 1978, pp. 14-16). These authors are nonetheless quick to underscore the tremendous value of impact evaluation. Rossi and his group (1979) consider impact evaluation synonymous with causality determination.

The basic aim of impact assessment is to estimate the net effects or net outcomes of an intervention. Net effects or net outcomes are those results attributable to the intervention, free and clear of the effects of other elements present in the situation under evaluation. (p. 163).

There is an abundance of designs for evaluation as discussed in Chapter 2. The purpose of this checklist item is simply to assure that a technique to assess impact is identified in the program proposal. The question as to how well the evaluation method chosen will affirm causality and serve its purpose must be considered from several different angles, such as internal validity and selection methods, which are themselves discrete checklist items.

Rating scale--2 (optimum) = proposal clearly identifies an impact evaluation method. 1 (partially acceptable) = proposal alludes to an evaluation plan that will not directly indicate the effectiveness of the program. 0 (inadequate) = no plan for assessing overall program effectiveness is evident in the proposal.

5. Evaluation design. Description--the purpose of this checklist item is to expand on the previous item, impact evaluation plan, and to assure a thorough description of the evaluation design to be used. The underlying rationale for choosing the design should be included (ERS, 1980). The description should include the following information:

- a. Will one or more comparison groups be used?
- b. Will the comparison group(s) receive a variation of the program (treatment) or be excluded from the program, which is referred to as a no-treatment control group?
- c. How will the members of participant group(s) and no-treatment control group (if applicable), be selected from the entire target population?

d. How will the program (treatment) be distributed?
(three possibilities):

(1) Treatment will be provided to everyone who volunteers from the target population.

(2) Prospective participants will choose which treatment group (or control group) they prefer.

(3) Program staff will decide who will receive which treatment (or who will not) according to some preestablished system.

e. Will some baseline measure of performance (or pre-test) be determined?

f. How long after exposure to the program will each participant (and control if applicable) be evaluated for impact? Will the time interval between treatment exposure and impact evaluation be approximately the same for all participants?

g. How will the overall effectiveness be determined, e.g., comparison of treatment group results to control group results, comparison of one treatment group's results to the other treatment group's results, comparison of impact evaluation results (posttests) to baseline measures (pre-tests), comparison of participant's results to preestablished standards, or expert opinion (panel discussion, advocate vs. adversary debate, or an impartial judge).

For the purposes of this guideline, the plethora of evaluation designs and models are grouped into five general classes.

The particular strengths and weaknesses of each of the following design categories will be elaborated on under item 10, internal and external validity:

a. Case study--a single group is evaluated following exposure to the program. The results of such an evaluation are implicitly compared with other events casually observed and remembered (Campbell and Stanley, 1963, p. 6). Case study designs have been referred to as "preexperimental" (Campbell and Stanley, 1963; Fink and Kosecoff, 1978, p. 15) because they can be used to explore relationships; the discovery of relationships should be the stimulus for more rigorously controlled designs which are required to establish causality. For example, suppose a training program staff conducts an onsite inspection of laboratories that participated in a field trial of their program. They find very few deficiencies and would like to believe that their program was responsible for the high degree of compliance to standards. The staff consults with the state Medicare Surveyors. The Surveyors seem to remember many more deficiencies the last time they officially surveyed the laboratories which was before the training program was disseminated. There appears to be some relationship between training and fewer inspection deficiencies. However, in order to establish that training caused fewer deficiencies, several rival explanations would have to be ruled out. The program staff decides to develop a large scale training

program and designs an evaluation which will allow causal inferences about the effects of training on inspection deficiencies.

b. Before and after design--also known as the one-group pretest-posttest design (Taylor-Fitz-Gibbon, Lyons-Morris, 1978a; Campbell and Stanley, 1963). Some measure of performance is taken before the program is distributed to the participants. After all participants have received the program, they are retested, usually with the same or an equivalent measurement instrument. Campbell and Stanley (1963) also dubbed this design preexperimental since the same uses and cautions apply here as for case studies. There are many variables which could explain differences between pre and post program performance and thus invalidate cause and effect conclusions.

c. Static group comparison--in this design, a group of program participants is evaluated along with a group of non-participants. The two groups may have some similar characteristics, but they are not equivalent. Therefore, cause and effect conclusions are once again precluded. For example, a group of State regulated laboratories could be compared to a group of voluntarily accredited laboratories to see whether the voluntary program is more effective in assuring laboratory quality (as measured by proficiency tests) than the regulatory program. Even if the laboratories in both groups were the same size, geographic

location, and served the same type of patient population, other variables besides accreditation could be the cause of any performance differences between the groups. Perhaps laboratories seeking voluntary accreditation would perform better because they are generally more conscientious. Accreditation, in this instance, would be an extra status symbol, not the means to achieve quality.

Campbell and Stanley (1963) characterized the static group comparison as an ex-post-facto design, which was classified as a non-experimental design in Chapter 2. However, static group comparisons can also be considered pre-experimental along with case studies and before and after designs. The distinction between Campbell and Stanley's term preexperimental and the term nonexperimental introduced earlier in this thesis is nominal. Both terms are meant to convey the clear inferiority of uncontrolled designs in establishing causation compared to true experimental designs. This characteristic does not imply that the information collected from a nonexperimental design is totally worthless, only that causal inferences derived from the information are invalid.

d. Time series design--also referred to as longitudinal evaluation (Fink and Kosecoff, 1978; Taylor-Fitz-Gibbon and Lyons-Morris, 1978a). A baseline of performance is established and compared to post program performance sometime later. Performance should be measured at least three times before program implementation and three times after program

completion (Taylor-Fitz-Gibbon and Lyons-Morris, 1978a) to determine any discontinuity that would be attributable to the program. Time series designs are especially useful for determining whether program effects are long lasting or only ephemeral. Conclusions drawn from time series designs are not immune to opposition. Campbell and Stanley (1963) classified time series designs as quasi-experimental to elevate their status somewhat above preexperimental (or nonexperimental) designs. However, there are still one or two sources of invalidity threatening cause-effect conclusions from time series designs. For example, the finding that laboratory proficiency test scores increase substantially over the years (La Motte, 1977) is indicative of a relationship between years of proficiency test participation and performance improvement. However, to unequivocally conclude that prolonged enrollment in proficiency test programs causes performance to improve, one would have to employ a much more rigorous evaluation design to rule out rival explanations of improved performance, such as technological advances, better trained personnel, and the like.

c. Maneuverable group comparison--comparison group designs can be either quasi-experimental or true-experimental designs (Fink and Kosecoff, 1978). If participants are assigned to treatment and control groups in such a way that the two groups may not be equivalent, the design is quasi-experimental. An example would be where two equally

attractive variations of the program would be offered and participants would be allowed to select their preference. Another example would be as follows: Suppose a technical consultation program was developed to improve laboratory proficiency test scores. Some laboratories were enrolled in a proficiency test program through their professional organization whereas the other laboratories were enrolled in the State's proficiency test program. The technical consultation program staff decided to assign the one group of laboratories to the technical consultation program and to designate the other group as the control. They made the assignment randomly. They could then examine proficiency test scores before and after the program for each group, and compare the treated group's gains to the control group's gains. An even greater amount of validity could be assured if they would examine proficiency test scores of both groups over a similar time period prior to program implementation. They could then compare the overall net gain for each group.

A true experimental design requires that participants be randomly selected from the target population so that a representative sample of participants is assured. The participants must then be randomly assigned to treatment or control groups. This process affords maximum validity to deriving causal inferences from evaluation results and to generalizing findings to the total target population. Whether or not a pretest is used, random assignment can be

assumed to guarantee the pretest equivalence of treatment and control groups (Campbell and Stanley, 1963).

Rating scale--4 (maximally interpretable) = maneuverable group comparison design is described in the proposal for the impact evaluation. 3 (moderately interpretable) = a time series design is described. 2 (possibly interpretable) = a static group comparison is described. 1 (borderline uninterpretable) = a before and after design is described. 0 (uninterpretable) = a case study is described, or else no particular design is evident in the proposal.

Interpretable refers to cause and effect inferences. The rationale for this rating scale stems from Campbell and Stanley's discussion on sources of invalidity of designs (1963). Item #10 will describe specific strengths and weaknesses of each design category. Within a particular category, many design arrangements are possible. Categories can even be combined to yield novel designs and to evaluate different elements of a total program.

6. Inferences intended. Description--in order to match the design to the desired inference (or vice versa), the conclusions that the program hopes to derive must be clearly stated (ERS, 1980). The inference should be supported by the design. If the desired inference is to determine the effectiveness of a laboratory improvement program in upgrading laboratory performance, the design must be a maneuverable group comparison in order to warrant the cause and effect inference. If one of the other designs is to be used, the wording of the inference must reflect

restraint. For example, for a nonexperimental design, the inference might be stated as follows: to explore the possible relationships between laboratory performance and participation in a training program. For a time series design, the inference could be stated as follows: to determine whether the relatively stable fluctuations in laboratory performance shift substantially following participation in a technical consultation program. Whether or not the inference is supported by the design is a significant issue considered in item 10, internal and external validity. This item, inferences intended, is primarily concerned with whether the inference is clearly spelled out in the proposal. Even so, if this checklist item is to be used to aid future proposal writers, the wording of the inference statement should be carefully reviewed at this point so that later revisions will not be necessary.

Rating scale--2 (optimum) = proposal clearly states inference intended. 1 (partially acceptable) = the inference can be partially deduced from the proposal's goals and purposes or other discussion of the impact evaluation plan, but it is not specifically stated anywhere in the proposal. 0 (unacceptable) = proposal does not provide any reference to the conclusions to be drawn from the evaluation results.

7. Statistical tests. Description--the field of statistics has traditionally been divided into two categories:

- a. Descriptive methods to reduce data into meaningful values including measures of central tendency such as the

mean and median, and measures of dispersion such as the standard deviation.

b. Inferential techniques to answer questions about differences between two samples or generalization from a sample to the population (Bartz, 1976, Chaps. 1, 8 and 9). Inferential statistics include t tests, analysis of variance, and chi square tests.

A third category sometimes included under descriptive statistics has to do with methods for examining relationships between variables. Common measures of relation are correlation and linear regression.

If a program intended to improve laboratory performance, it is not enough to say the performance of the treated group is three percent better than it was before treatment or than the performance of the control group. Although this statement may be based on a mathematical calculation, it does not provide insight into the statistical or medical significance of three percent. The goal of the program and the inferences to be drawn require that inferential statistics be used in this instance. Scriven once said such tests require no great sophistication (1974, p. 17); however, in a later work, he qualified his position somewhat. "Statistics (in evaluation) are already pretty sophisticated, although their selection and interpretation still require a good deal of judgment" (1976, p. 134). Whether or not the program staff has a firm grasp of statistical methods, Stufflebeam recommends consulting with a competent statistician

to verify that the data analysis plan is appropriate and sufficient (1978, p. 36). The ERS standards specify that the statistical analyses be matched to the evaluation design (1980, p. 17). Fink and Kosecoff (1978, pp. 47-70) and Taylor-Fitz-Gibbon and Lyons-Morris (1978b) provide brief but comprehensive descriptions of how analyses can be done with or without computer processing. These works are specifically written for program evaluators who are new to the role.

Rating scale--2 (optimum) = proposal discusses intention to apply inferential statistics and names at least one statistical analysis to be performed. Statistical tests to examine relationships or associations between variables qualify as inferential if they are more appropriate to the evaluation design and intended inferences. 1 (partially acceptable) = proposal mentions at least the descriptive statistics to be used to illustrate possible program effects such as mean and standard deviation of treatment and control groups' proficiency test scores. 0 (inadequate) = proposal does not mention any statistical analyses or methods to reduce impact evaluation data into a quantitative form.

8. Unit of analysis. Description--along with the evaluation design, inferences intended, and statistical analyses, the unit of analysis should be contemplated in advance of the program implementation (ERS, 1980). The contractor must determine whether the individual or the organization is the more appropriate unit of analysis. For example, a training program may decide the

individual is the basic unit to be measured. The individuals are then randomly assigned to receive two different treatments. Consider what would happen if technologists from the same laboratory are assigned to receive different treatment. Contamination is likely to occur (Cook and Campbell, 1976, p. 302) and obscure the differential effects of the treatments if any would truly exist. The solution would be to either limit the training to one individual per laboratory or to consider the entire laboratory as the unit of analysis and randomize laboratories to treatment groups. If several individuals from the same laboratory attend, their results on the performance measure or dependent variable would be averaged. The decision about which unit to measure is an important one. The contractor should consult with the funding agent to determine whether a laboratory improvement program should focus on the entire organization and seek change in the laboratory's structure, environment, and personnel, or just in the individual's characteristics.

Rating scale--4 (optimum) = proposal considers both the laboratory and the individual as the units of analysis. The methods for measuring laboratory performance, e.g., proficiency testing and onsite inspections, and individual laboratory workers' performance or knowledge are described. 3 (partially preferable) = proposal clearly designates the laboratory as the unit of analysis. 2 (marginally satisfactory) = proposal alludes to the laboratory as the unit of analysis but no clear decision is evident between measures of laboratory performance and

measure of individual performance. 1 (borderline inadequate) = proposal indicates a preference for the individual, rather than the laboratory as the unit of analysis. Measures of laboratory performance will clearly not be included. 0 (inadequate) = proposal does not even imply whether the laboratory or the individual worker will be measured.

The rationale for this rating scale is derived from CDC's RFP's (1979). The weightings could easily be transposed if the funding agent changes the stipulations or the major purpose they envision for the programs.

9. Method of selection and assignment. Description

--Stufflebeam (1978) suggested this standard which means that the method used to select a sample (of laboratories) must be described. The sample selected should be representative of the entire target population. If a program cannot be offered to everyone in the target population, this is a situation especially conducive to random selection which assures representativeness (Cook and Campbell, 1976). Random means that every member of the target population has an equal likelihood of being selected. Random samples are best drawn from the population using a table of random numbers. These are available in most introductory statistics texts (Bartz, 1976, pp. 388-391). Random, in this context, does not mean haphazard or capricious (Rossi et al., 1979, p. 183).

Random sampling must be distinguished from random assignment. Random sampling assures adequate representation of the target

population in the entire study group. Random assignment ensures that the treatment and control groups are equivalent (Borg and Gall, 1979, p. 193). Individuals in the treatment group will not be exactly like individuals comprising the control group, but if there are more than just a few individuals within each group (say 10 or more), individual differences will not exceed chance fluctuations that are to be expected (Rossi et al., 1979, p. 184). Differences between the composition of the two groups can be further reduced if obvious outliers (on pretest measures only) are eliminated, not necessarily from the treatment group, but at least from subsequent data analyses (Borg and Gall, 1979, p. 194). If two treatments (or one treatment and a control group) are to be compared, it is always preferable to randomly assign or randomize participants to type of treatment rather than allow them to select themselves into a treatment group (Boruch, 1976; Campbell and Stanley, 1963; Taylor-Fitz-Gibbon, and Lyons-Morris, 1979, pp. 24-25). Again, a table of random numbers should be used. If the program can only recruit volunteers, some options are still available for randomization as discussed in Chapter 2.

It should be clear that random assignment does not assure adequate representation of the target population. To maximize internal and external validity, the program should consider randomly selecting laboratories (or individuals) from the entire target population and randomly assigning those selected to comparison groups. Of course, random sampling is unnecessary

if all members of the target population will be included in the evaluation.

If the contractor or program staff selects the individuals (or laboratories) to receive treatment based on some contractor-set criteria, the results are generalizable only to other laboratories under similar circumstances, and the treated individuals should not be compared to any group of nonparticipants. Whatever the selection system, it should be explained completely and justified (ERS, 1980).

Rating scale--4 (optimum) = proposal describes plan to randomly select and randomly assign (randomize) units (laboratories or individuals) to treatment and control groups. 3 (partially preferable) = proposal describes plan to select a representative sample using a systematic or stratified random sampling process. No mention is made of assignment to treatment group methods or assignment to treatment group will be made in a nonrandom fashion. 2 (possibly sufficient) = proposal describes plan for program staff to select or assign participants based on preset criteria to enroll participants who demonstrate need. 1 (marginal) = proposal describes intention to allow volunteers to self-select into treatment groups. Program staff will prioritize those to receive treatment based on preset criteria to enroll those who most need the program if more individuals volunteer than can be accommodated. 0 (undesirable) = proposal describes plan to recruit and accept any and all volunteers to receive treatment(s). No criteria are established for assignment

to treatment group or preferential enrollment. No plan is described to assure adequate representation of the target population.

10. Internal and external validity. Description--this is the crux of the evaluation design issue. Different evaluation designs afford varying levels of validity (Campbell and Stanley, 1963). The concepts of internal and external validity were introduced and defined in Chapter 2. They require further elucidation to justify their inclusion as a separate checklist item. Although there are standard works on evaluation designs (Campbell and Stanley, 1963; Cook and Campbell, 1976) which detail the threats to validity found in customary designs, it is inevitable that designs assume their own unique identity as a result of logistical and political constraints. Each individual design must be closely scrutinized to uncover its own peculiar vulnerabilities before it is implemented. If the design is not sound at the planning stage, it is sure to degenerate during its execution when the unpredictable ways of reality take their toll. To prevent poorly conceived designs, the plans should be checked against each of the common threats to validity. This should be done even if a well known commonly used design is planned. The usual threats to internal validity (robustness of causal inference) are as follows (Fink and Kosecoff, 1978, pp. 13-14):

- a. History--the effects of changes in the environment that occur simultaneously with the program being evaluated.
- b. Maturation--the effects of physiological and

psychological changes in participants brought about by the ordinary passage of time.

c. Testing--the effects that the experience of test taking alone has on later performance measures. If pretests are given, they can cause improved performance on the posttests without any other treatment.

d. Instrumentation--changes in the way measurements are taken and scored which may cause spurious post treatment results.

e. Statistical regression--the fact that those who achieve very high scores tend to score lower the next time they are tested and those who score very low the first time do better the second time. When participants are selected because of their extreme performance levels, the changes observed on the posttest must be considered in light of statistical regression.

f. Selection--the method of assignment to treatment group does not assure equivalence (nonrandom) and results in specious findings.

g. Mortality--also known as participant attrition. If the treatment groups are not equivalent, drop out rates of participants in each group may differ and render the results uninterpretable.

h. Selection interaction with maturation, testing, or history--a difference in performance between treatment groups which can be explained by characteristics of one

group being different from the other group. These characteristics formed the basis for selection of the two groups and thus the characteristics are said to interact with the way they were selected; the interaction confounds the proper interpretation of the group differences on the post-test (Campbell and Stanley, 1963, p. 48).

The threats to external validity (generalizability) are as follows (Fink and Kosecoff, 1978, p. 14):

a. Reactive effects of testing--participants have been sensitized to the pretest. The effects of the program attributable to the treatment alone are unknown.

b. Hawthorne effect--the novelty of the program causes participants to change their performance. This is also known as the placebo effect in medical research (Borg and Gall, 1979, p. 528).

c. Interaction of selection and treatment--only certain types of individuals volunteer to receive the treatment. This prevents generalization of their results to the entire target population.

d. Interaction of history and treatment--the particular time period in which the treatment is distributed is responsible for its effectiveness, not the treatment alone.

e. Experimenter effect--the treatment is effective (or ineffective) because of the type of person administering it, not because of the treatment's own merits (Borg and Gall, 1979, p. 527).

f. Multiple treatment interaction--changes observed are due to participants receiving several different programs simultaneously. The effects of any one treatment alone are indeterminate (Fink and Kosecoff, 1978, p. 14).

g. Time of measurement effects--measurement of impact is done too early or too late to determine the overall effects of the program (Borg and Gall, 1979, p. 527).

Referring back to item 5, evaluation design, the general weaknesses of each of the five designs described can be pinpointed. The internal validity of a case study is threatened by history, maturation, selection and mortality. No variables are controlled. The external validity is vulnerable to interaction of selection and treatment and time of measurement effects.

Before and after designs are susceptible to history, maturation, testing, instrumentation and interaction of selection and maturation, etc. under internal validity; and reactive effects of testing, and interaction of selection and treatment under external validity. Selection and maturation are controlled.

The static group comparison is threatened by selection, mortality, and selection interactions with maturation, etc. Its external validity is threatened by interaction of selection and treatment. History, testing, instrumentation and regression are controlled in this design.

The time series design is usually only vulnerable to history in terms of internal validity, the other factors are controlled, as long as the instrumentation remains consistent. This design

is vulnerable to the reactive effects of testing under external validity.

The quasi-experimental (maneuverable group) comparison design is vulnerable to interactions of selection and maturation, etc. All other variables are controlled, provided the comparison groups are similar enough to prevent statistical regression from confounding true effects. The external validity of this design is threatened by the reactive effects of testing.

Lastly, the true-experimental (maneuverable) comparison group designs control for all the threats to internal validity. Some true experimental designs are threatened by reactive effects of testing, others control for this aspect of external validity (Campbell and Stanley, 1963).

Smith and Glass (1977) recently developed a rating system for the validity of evaluation designs. Designs that rate high on the scale exhibit low mortality and minimize threats to validity. The requirement for a design to rate high is that it be based on randomization. A design is rated medium if it carries more than one threat to internal validity (Smith and Glass, 1977, p. 755). Finally, the designs that fail to match or equate different treatment groups, or lack baseline (time-series) data, are rated low.

Rating scale--2 (optimum) = proposal describes plan to use a true experimental design with maximum internal validity.

1 (partially acceptable) = proposal describes plan to use a design with only one or two possible threats to internal validity

and only one clear threat to external validity. 0 (inadequate) = proposal describes plan to use a design with three or more threats to internal validity and one or more threats to external validity.

11. Unbiased evaluator. Description--the underlying concept here has been well articulated by Scriven (1976) who argued that the best way to keep an evaluation from becoming biased is to establish and periodically reestablish the independence of the evaluator. An impartial evaluator should at least be invited to consult on the project if not conduct the entire evaluation. Stufflebeam (1978) called for objective evaluators, in his evaluation standards, and urged that personal feelings and prejudices not be allowed to distort objective evaluation.

Rating scale--2 (optimum) = proposal describes plan to appoint a neutral and credible evaluator who will provide assistance and some assurance that independent judgment on program impact will be possible. 1 (partially acceptable) = proposal shows signs of the potential for unbiased judgment by describing plans to consult an outside evaluator for assistance with some (but not all) aspects of the evaluation, such as statistical advice; or an individual from within the contractor's organization will be asked to help conduct the evaluation for the purpose of adding credibility to judgments about impact. 0 (inadequate) = no possibility for independent judgment from an external or internal evaluator is apparent in the proposal. The program staff will evaluate all of their own efforts.

12. Evaluation meets audience objectives. Description --both Stufflebeam's (1978) and the ERS's (1980) evaluation standards discussed the importance of identifying the audience's (or funding agency's in this context) needs in the overall evaluation system. Lyons-Morris and Taylor-Fitz Gibbon (1978) assert that the determination of what the commissioner of an evaluation really wants from the evaluation is step number one for any program. The key questions are whether the audience is more interested in implementation or outcomes; whether the evaluation will have an opportunity to set up an experimental or quasi-experimental design to maximize the validity of the findings, or be restricted to the approximate methods for impact assessment as described by Rossi et al. (1979, pp. 227-243).

Rating scale--2 (optimum) = proposal describes evaluation plan which will attempt to answer the question: How effective is the program in bringing about improved laboratory performance. Some measure of laboratory performance must be described (not attitude change). 1 (partially acceptable) = proposal describes evaluation plan which partially or indirectly addresses the funding agent's question, e.g., participant's self-report will be used rather than performance measures, or the plan is not clear whether knowledge or performance will be measured. 0 (inadequate) = proposal does not describe any plan to address the funding agent's evaluation question.

13. Evaluation meets program objectives. Description--the specific objectives stated in the proposal are linked directly

to the evaluation plan. For example, if a program states an intention is to improve attitudes toward preventive maintenance of equipment, the evaluation plan must somehow seek to measure these attitudes in addition to performance. The program purpose, objectives, audience needs, and inferences to be drawn must all be consistently integrated into the evaluation plan. An evaluation measurement must be described for every objective stated in the proposal.

Rating scale--2 (optimum) = proposal delineates evaluation and measurement method for every goal and objective stated.
1 (partially acceptable) = proposal provides incomplete evaluation and measurement plans. Some objectives are not followed by the means for measuring achievement. 0 (inadequate) = proposal's description of measurement methods does not relate to the objectives stated, there is an obvious inconsistency between the goals and the evaluation and measurement methods.

14. Plan to measure unintended outcomes. Description--these are the possible side effects of a program (Scriven, 1974). They can be either favorable or undesirable. The object here is to assure a systematic collection of this kind of pertinent data. For example, the program proposals could plan to gather this information through surveys of clinicians, interviews with supervisors, and records of laboratory's changes in methods. Although the term "unintended" may be interpreted as undesirable, Scriven (1974, p. 35) contends that these effects might well be the crucial achievement. The full benefits of detecting unintended

outcomes require a good eye for the future and some healthy imagination on the part of the evaluator who must anticipate the unexpected.

Rating scale--2 (optimum) = proposal describes at least one method to detect unintended outcomes that will indicate program impact. 1 (partially acceptable) = proposal describes a method which may uncover unintended outcomes for use in the formative evaluation, but not the impact evaluation; or proposal mentions a method that might be considered, but no clear intention to use the method is evident. 0 (inadequate) = proposal does not describe any additional measurement methods or plans beyond those that meet the major program goal (or audience's goal).

15. Plan to evaluate long term effects. Description--this entails a plan to follow-up on the initial evaluation, to be able to measure the outcomes like general attitudes that take more time to surface (Scriven, 1974, p. 16; Campbell and Stanley, 1963). Some argue that long term effects can only be postulated by the evaluator (Shepherd, 1977). In eighteen months, it may be possible only to follow-up on the very first workshop given in the beginning of a training program. Sources to search for longitudinal data include proficiency testing, survey inspections, reports from manufacturer or drug company detail persons, and telephone or postcard surveys to former participants and their employers. Again, the creativity of the evaluator may turn up effects that would otherwise be lost. Extreme caution must be exercised in the interpretation of such data as other

intervening variables have surely entered the picture. Nevertheless, a longitudinal perspective will insure discovery of new questions (Stufflebeam, 1978).

Rating scale--2 (optimum) = proposal describes intention to measure long term effects with an identification of at least one measurement method. 1 (partially acceptable) = proposal implies intention or possibility to detect long term effects, but no measurement methods are described. 0 (inadequate) = no possibilities for measurement of the program's long term effects are stated or implied in the proposal.

To summarize the preceding discussion, the category of structure consisted of 16 items which are interrelated because of their collective focus on the organization of evaluation. As the context category depicts the overall appearance of the program and its setting, the structure category represents the layout of the evaluation with all the features vital to the conduct of laboratory improvement program evaluation. One ramification of the total evaluation picture that tacitly runs through most of the structural components is the third technical category--instrumentation.

C. Instrumentation. This category encompasses seven evaluation elements, all relating to the collection of evaluation information. Returning to the analogy of constructing a building (applied to shaping a program evaluation), once the completed framework is resting on the foundation, the final step is to fill in internal details. The instrumentation items are the final details necessary to "rough-in" the technical domain of evaluation planning, and make

resulting product recognizable as valid evaluative inquiry. Table 1 shows how the instrumentation items fit into the technical aspects of evaluation and within the comprehensive evaluation scheme.

1. Measurement methods. Description--measurement methods refer to a broad class of techniques available to the program evaluator. The ERS (1980) standards specify that measurement methods be identified. A contractor may use general terms to convey the methods, e.g., written tests, onsite proficiency testing, observation using a performance checklist, etc. The essential point here is to clearly state the intentions and include the rationale for the use of the methods. Campbell and Stanley (1963) advocate the use of multiple methods to increase validity. The methods should be cross checked with the specific program objectives, audience objectives, evaluation design and inferences intended. If there is any inconsistency, this will be the last and best opportunity to weed it out.

Rating scale--2 (optimum) = measurement methods are clearly identified and justified in the proposal. 1 (partially acceptable) = measurement methods are implied in objectives but not clearly identified, or methods are identified but no rationale is offered. 0 (inadequate) = no measurement methods can be discerned in the proposal.

2. Identification of instruments. Description--the origins of this item can be found again in Stufflebeam (1978) and the ERS (1980). The particular instruments to be used should be explicitly defined, as much as possible at the proposal stage.

If multiple choice tests are to be used, a few sample questions would represent the appropriate level of specificity. If proficiency testing will be the primary instrument, the contents should be described and justified. Even questionnaires and telephone survey scripts should be roughly conceptualized. The instruments must also be integrated into the evaluation system so that they match goals, designs, inferences and planned measurement methods. Each method described in the previous item must have at least one instrument identified for it.

Rating scale--2 (optimum) = instruments are specifically identified and at least one instrument is identified and described (with examples) for each method listed under the previous checklist item. 1 (partially acceptable) = some instruments are identified but not all methods listed have a corresponding measurement instrument described. 0 (inadequate) = proposal does not list or describe any specific measurement instruments.

3. Estimation of validity. Description--the subcategories of instrument validity were discussed in Chapter 2, i.e., content, criterion-related and construct validity. Content validity can be increased in written tests if educational objectives and content topics are both considered when test items are developed (Gronlund, 1976). A contract proposal will usually only roughly estimate validity or explain how more accurate estimates will be derived. Gronlund's (1976) text provides the necessary techniques for determination of validity. Especially important in laboratory evaluation is the construct

validity of proficiency testing and onsite inspections (Peddecord, 1978). More attention to this issue on the part of program developers could contribute a great deal to better and more meaningful measures of laboratory performance.

Rating scale--2 (optimum) = some estimation of the validity of each measurement instrument is provided. The estimation does not have to be mathematically derived as long as some plan to pilot test the instrument is included in the proposal. A qualitative estimation would suffice until the calculations could be determined. 1 (partially acceptable) = proposal indicates some awareness of the need for validity estimates by describing a general intention and a rough plan to maximize instrument accuracy or precision. At least expert panel review would be described in the plan. 0 (inadequate) = proposal does not even mention validity of instruments as a concern.

4. Estimation of reliability. Description--reliability refers to the repeatability of the test results. A 100 percent reliable test will give the same results every time it is administered. However, reliability should not be confused with validity; the same way precision should not be confused with accuracy in clinical laboratory results. The methods to calculate reliability are fairly straightforward and again well described in Gronlund (1976). The practice of sending proficiency test specimens to reference laboratories and performing extensive quality control is related to reliability. But the full spectrum of reliability has to do with how the tests

are scored as well. Different scorers should derive the same results. A reliable checklist to be used for observational purposes will yield the same participant scores even if several different people administer the instrument. This describes the concept of interrater reliability mentioned in Chapter 2.

Rating scale--2 (optimum) = proposal provides qualitative estimation of instrument reliability or describes plan to furnish quantitative estimation of reliability before instruments are applied to the entire participant group (pilot test).
1 (partially acceptable) = proposal does not explicitly provide plans to assess reliability, however, the potential for determining reliability is implicit in its description of measurement methods, e.g., more than one observer/rater will be used, institutional quality control records will be reviewed periodically, etc.
0 (inadequate) = proposal exhibits no concern for reliability. No explicit plans are described and the potential for reliability estimates cannot be inferred.

5. Judgment criteria/standards. Description--judgment standards apply to the scoring of tests, as in proficiency tests, where some standardized system is required to yield reliable data (Forney et al., 1978). If a training program plans to use criterion referenced tests (Fisk and Kosecoff, 1978, p. 33; Gronlund, 1976, p. 19) again the judgment standards must be spelled out, e.g., 95 percent correct will be considered sufficient mastery of the training objective. Standards should be adopted (or adapted) from reputable sources and the references

must be cited. Stufflebeam (1978) also uses the term judgment standards to connote the overall standards that will be used to judge the program's effectiveness. For example, a program might state: Technical consultation will be considered effective if the pre and posttest gain of the treated group is (statistically) significantly different from the control group's pre and posttest gain.

Rating scale--2 (optimum) = proposal clearly describes judgment criteria and standards as they apply to particular measurements and data analyses to be used in the evaluation, i.e., wherever grades are to be assigned, ratings are to be made, or participants are expected to achieve certain performance levels before the program is considered successful (including statistically significant pre posttest differences or treatment-control group differences in performance). References are cited or other justification is provided. 1 (partially acceptable) = proposal describes judgment criteria for measurements, but not for program effectiveness; or standards for judging program effectiveness are provided, but not measurements. Incomplete judgment criteria, unclear judgment standards and nonreferenced or justified judgment standards would also fit this category. 0 (inadequate) = no judgment criteria for measurements or judgment standards for determining program effectiveness are alluded to in the proposal.

6. Reactivity of measurement. Description--the Smith and Glass (1977) paper was the original source for this item.

Posavac (1979) redefined Smith's and Glass's scale, and made the categories more adaptable for laboratory improvement program evaluation. A measure's reactivity is related to how easy it is to fake a response. If a test is artificial, participants' scores will bear little relationship to the measures of typical performance desired. The following levels of reactivity apply to this checklist item for laboratory evaluation (low is desirable, high is undesirable).

a. Low--blind record audit of proficiency test data, proficiency test scores from a program outside the purview of the contractor, blind proficiency testing.

b. Inspection or accreditation visit from some agency other than the contractor's, and "blind" to experimental conditions, or at least impartial.

c. Specially constructed cognitive tests, PT scores from within the contractor's organization.

d. Specially constructed opinion/attitude surveys; self-report measures, supervisor's ratings (non-blind).

e. High--non-blind ratings of an observer, inspections initiated by the contractor.

Each measurement method to be included in the evaluation should be evaluated for its reactivity. If one or two planned measurements rate high in reactivity (undesirable), the program staff should not discard the measurement plans, but simply make certain that at least one planned measurement rates low in reactivity. The more measures taken, the more dependable the

data (Webb et al., 1966). Therefore, even reactive measures can yield important information when examined in light of more unobtrusive measures.

Rating scale--4 (optimum) = proposal describes plans to include multiple measures (two minimum) at least one of which rates a low of one in reactivity. 3 (partially preferable) = proposal describes plan to include multiple measures (two minimum) at least one of which rates a two on reactivity. 2 (minimally satisfactory) = proposal describes plan to include multiple measures (two minimum) at least one of which rates a three on reactivity. 1 (borderline inadequate) = proposal describes plan to include multiple measures (two minimum) at least one of which rates a four in reactivity; or, proposal plans to use only one single measure which rates a three on reactivity. 0 (inadequate) = proposal describes plan to include only one single measure which rates a four or five on reactivity.

7. Data management. Description--several sources stress the absolute necessity of maintaining adequate, accurate records (Hamilton, J., 1977; Stufflebeam, 1978; ERS, 1980). The plan for coding, storing, and retrieving data should be developed well in advance of program implementation. The use of a computer is recommended if sophisticated analyses of the data will later be desired.

Rating scale--2 (optimum) = proposal describes the data management plan including how information will be recorded, coded, sorted, and organized into data files. If data will be

entered into a computer, the above details can be inferred; however, the proposal should make it clear that the information will be readily retrievable and decipherable. 1 (partially acceptable) = proposal mentions that a filing and data management system will be maintained, but fails to describe the procedures to be followed. 0 (inadequate) = proposal provides no plan for or apparent consideration of data management.

In summary, the instrumentation category subsumed seven individual items all relating to data collection issues. The term instrumentation refers to the application of measurement tools to document behavioral attributes or laboratory conditions and policies that are expected to vary as a result of laboratory improvement efforts. The specific items outline enough of the process to assure that data collected are representative in their scope, directly indicative of relevant laboratory performance characteristics, and amenable to statistical analyses.

To conclude this section, thirty-one checklist items were categorized under the headings of Context, Structure and Instrumentation. The overall scheme in Table 1 was elaborated on by defining each category and describing in detail the checklist items subsumed under the category.

The context of the program refers to the events and conditions leading up to the conception of the program. As such, the context builds the foundation for the evaluation. The structure of the evaluation is built upon the underlying context and provides the framework for valid evaluative inquiry. Instrumentation is the

extension of the evaluation framework. The instrumentation elements are the least abstract in the domain of technical quality. Thus, the complete spectrum of the technical aspects of evaluation as they relate to program proposals has been presented, from the conceptual to the applied level of analysis. Technical aspects deal with the organization of valid evaluation grounded in evaluation and measurement theory. The entire scheme of program evaluation encompasses conceptual, technical, ethical, economic, and effectual categories. Thus it can be seen that the thirty-one guidelines described in this chapter cover only one-fifth of the total field of program evaluation.

Analysis and Results

Twenty-three laboratory improvement program proposals, which had been awarded contracts by CDC, were reviewed. The review can be considered a formative meta analysis. Stufflebeam (1978) conceptualized formative meta evaluation as a constructive enterprise that aids evaluators in conceiving, planning, conducting, interpreting, and reporting their studies (p. 23).

Formative meta evaluation has its foundation in evaluation guidelines . . . and assesses the extent that the plan . . . of an evaluation study measures up to guidelines which, if followed, will result in sound evaluation studies. (Stufflebeam, 1978, p. 24).

The analytical review discussed in this section typifies the formative meta evaluation envisioned by Stufflebeam. The checklist of the guidelines presented in the previous section was used as the measurement criterion for evaluating the technical integrity of the

twenty-three contract proposals.

The ensuing report of the findings is divided into the following subsections: (1) descriptive data, (2) inferential statistics, (3) relationships and (4) conclusions. The descriptive data subsection lays the groundwork for the three hypotheses tested under the inferential statistics subsection, and the two hypotheses addressed in the relationships subsection.

Descriptive Data

The results are presented in a data matrix (see Table 2). The matrix represents 713 independent judgments which were assigned ordinal ratings consistent with the rating scales introduced in the previous section. The salient aspects of Table 2 are condensed in the paragraphs to follow within this subsection.

Obvious strengths and many weaknesses in the technical quality of the proposals were revealed. The estimable characteristics will be discussed here, followed by a close examination of the inadequacies. In the majority of cases, the purpose was clearly defined, the target setting was described, the program approach was innovative, an impact evaluation plan was identified, and the evaluation appeared to address the audience's (CDC's) objectives. Every single proposal clearly indicated at least one measurement method to assess the program's effects. Particular strengths were as follows:

1. The majority of the proficiency test proposals rated high on description of a replicable, exportable program; identification of instruments, estimation of reliability; description of

Table 2

Ratings of the Evaluation Characteristics of Twenty-Three Laboratory Improvement Program Proposals

Program #	Proposal Type ^a	CONTEXT ITEMS									STRUCTURE ITEMS											INSTRUMENTATION ITEMS			Total Scores	Amount of Funds Awarded in \$	Prior Contract Experience								
		Clear Need	Defined Purpose	Description of Target Setting	Description of Target Population	Program Approach	Plan for Cooperation/PR	Assessment Plan-manifest Needs	Assessment Plan-perceived Needs	Replicable Exportable Program	Formative Evaluation Plan	Trial Run	Program Monitoring Plan	Impact Evaluation Plan	Evaluation Design ^b	Inferences Intended	Statistical Tests	Unit of Analysis ^b	Method of Selection/ Assignment ^b	Internal and External Validity	Unbiased Evaluator	Evaluation Meets Audience Objectives	Evaluation Meets Program Objectives	Provision to Measure Unintended Outcomes				Plan to Evaluate Long Term Effects	Measurement Methods	Identification of Instruments	Estimation of Validity	Estimation of Reliability	Judgment Criteria/Standards ^b	Reactivity of Measurement ^b	Data Management System
1	PT	0	0	2	0	0	0	2	0	2	0	2	2	2	0	2	3	0	0	0	1	0	0	1	2	2	2	0	2	2	1	2	26	48,006	2,2
2	PT	2	2	2	0	2	0	0	1	2	2	2	2	3	2	1	3	0	1	2	2	1	2	1	2	2	2	0	2	2	2	45	98,197	2,2	
3	PT	1	2	2	0	1	1	0	1	2	2	0	2	0	0	0	3	0	0	0	0	2	0	0	2	2	2	0	2	1	1	27	90,058	3,5	
4	PT	2	2	2	2	1	1	0	1	2	0	0	2	1	2	1	3	1	1	0	2	0	0	2	2	2	0	2	2	1	0	34	70,788	3,6	
5	PT	2	2	2	0	2	2	0	0	0	2	2	2	0	0	3	0	0	0	2	2	2	2	2	1	2	1	0	2	2	1	34.5	111,647	3,6	
6	PT	0	2	2	0	1	2	0	1	1	0	0	2	1	2	1	3	0	1	0	2	0	2	0	2	1	0	2	2	2	31	45,957	3,3		
7	TC	2	2	2	2	2	0	2	1	2	0	0	2	2	2	1	3	2	0	2	2	2	2	2	2	2	0	0	0	3	0	35	43,629	3,6	
8	TC	2	2	2	2	2	0	2	0	0	0	0	2	2	0	3	1	0	0	2	2	0	0	0	2	2	0	0	0	0	21	76,339	2,2		
9	TC	2	2	2	2	1	2	2	0	1	0	0	2	2	1	0	3	1	0	0	2	0	2	0	2	1	0	0	2	3	0	31	45,713	1,1	
10	TC	2	2	2	0	1	1	2	0	1	0	0	0	2	1	2	0	3	2	0	1	2	2	2	0	2	1	0	0	2	0	29	30,954	3,6	
11	TC	0	2	2	1	2	2	2	0	2	0	0	0	2	4	2	0	3	3	0	1	2	2	2	0	2	2	0	1	2	4	0	36	56,752	1,1
12	TC	2	1	2	2	2	1	0	1	2	0	0	2	2	4	2	0	3	3	2	0	2	0	2	0	2	1	0	0	3	0	34.5	22,613	2,2	
13	TC	1	2	0	2	2	1	2	1	1	0	0	0	2	3	2	0	3	2	0	1	2	0	0	0	2	1	0	0	4	0	28	32,791	2,3	
14	TR	1	2	2	2	2	2	2	1	2	2	1	0	2	4	2	0	4	0	0	0	1	2	2	0	2	2	0	2	0	1	0	38.5	99,226	3,6
15	TR	2	2	0	2	2	1	2	1	1	0	0	2	0	2	0	0	0	0	0	1	0	2	0	2	0	0	0	0	4	0	24	75,024	2,2	
16	TR	2	2	2	2	2	2	2	2	1	0	0	2	2	1	0	3	0	0	0	2	0	0	0	2	1	0	0	0	4	1	31	94,041	3,4	
17	TR	1	2	1	1	2	2	0	2	1	2	0	0	2	0	2	0	0	0	0	0	2	2	0	2	0	0	0	0	2	0	26	101,484	3,6	
18	TR	1	2	2	2	2	2	2	2	1	2	1	2	2	1	2	0	2	1	0	2	2	2	2	0	2	1	1	1	0	4	1	43	109,065	3,3
19	TR	1	2	2	2	2	2	2	2	2	2	0	2	0	0	0	1	0	0	0	1	0	1	0	2	1	0	0	0	4	0	28.5	71,364	1,1	
20	TR	1	2	2	1	0	1	2	1	1	2	0	2	2	1	2	0	4	1	0	0	2	2	2	0	2	1	0	1	0	4	0	34	65,238	1,1
21	TR	0	2	1	0	0	1	2	2	1	2	0	2	1	2	0	1	1	0	0	2	2	2	0	2	1	0	0	0	2	0	26.5	24,595	2,3	
22	TR	2	2	2	2	1	2	2	2	2	2	0	2	4	2	2	3	4	2	1	2	2	1	0	2	1	0	1	1	2	2	46.5	92,600	3,4	
23	TR	2	2	2	2	0	2	2	2	1	2	0	2	2	1	2	0	2	1	0	0	2	2	1	2	2	1	0	0	3	1	38.5	72,831	1,1	

^aPT = proficiency testing, TC = technical consultation, TR = training.

^bScales ranging from 0-4 were weighted as follows: 0=0, 1=0.5, 2=1.0, 3=1.5, 4=2.0

^cFirst number is the total number of years of contract experience, second number is the total number of contracts held.

judgment criteria or standards; and identification of a data management system.

2. The majority of the technical consultation proposals rated high on description of clear need, assessment of manifest needs, and inferences intended.

3. The majority of the training proposals rated high on description of the target population; plan for cooperation and public relations (PR); assessment of manifest needs; assessment of perceived needs; formative evaluation plan; inferences intended; evaluation plans consistent with program objectives; and provision to measure unintended outcomes.

Table 3 lists the major checklist deficiencies common to all twenty-three proposals and specific to each of the three different types of proposals, i.e., proficiency testing, technical consultation, and training. The data in Table 3 indicate that the greatest number of deficiencies for all three proposal types fell under the structure category. Table 3 shows 8 items deficient in this category across all three program types, and one additional deficiency in structure apparent in the technical consultation program proposals. One instrumentation item was deficient in all three proposal types; technical consultation proposals and training proposals each demonstrated two additional deficiencies in instrumentation. Context deficiencies were found only for proficiency test proposals (2) and technical consultation proposals (1) (see Table 3).

The ordinal ratings assigned to each of the 713 judgments transformed the data into a more quantitative form amenable to

Table 3
Major^a Checklist Item Deficiencies

	Category Label	Item Label	Percent of Proposals Inadequate
Deficiencies consistent across all three pro- posal types	Structure	Trial Run	83
	Structure	Program monitoring plan	61
	Structure	Evaluation Design ^b	74
	Structure	Statistical tests	74
	Structure	Method of selection/assignment ^c	74
	Structure	Internal and external validity	78
	Structure	Unbiased evaluator	70
	Structure	Plan to evaluate long term effects	74
	Instrumentation	Estimation of validity	91
Particular deficiencies in proficiency test proposals	Context	Description of target population	83
	Context	Assessment plan--manifest needs	83
Particular deficiencies in technical consultation pro- posals	Context	Assessment plan--perceived needs	71
	Structure	Formative evaluation plan	100
	Instrumentation	Estimation of reliability	86
	Instrumentation	Data management	100

Table 3 (continued)

	Category Label	Item Label	Percent of Proposals Inadequate
Particular deficiencies in training proposals	Instrumentation	Estimation of reliability	60
	Instrumentation	Judgment criteria/standards	90
	Instrumentation	Data management	60

^aMajor is defined as: greater than 60 percent of the proposals rated a 0 for the item.

^bProposals rated 0, 1, or 2 were considered inadequate.

^cProposals rated 0 or 1 were considered inadequate.

several statistical analyses. A single proposal could attain a maximum score of 63 points. The mean score achieved was 32.5 with a standard deviation of 6.6. The median score was 31. The highest checklist score was 46.5 and the lowest was 21.

Proposals of a certain type, i.e., proficiency testing, technical consultation and training, demonstrated several unique strengths and weaknesses; a pattern seemed to be in force. This finding instigated a search for possible determinants. The RFP's circulated by CDC (Center for Disease Control, 1979) were examined for possible overlap with some of the items comprising the checklist (see Tables 1 and 2). All three RFP's required the proposals to address specific items under the following headings (in the technical section): "Understanding the Problem," "Approach," "Personnel," "Facilities," and "Experience." Only the "Understanding the Problem" and "Approach" sections were relevant to this review. Table 4 shows the items listed by CDC and whether the RFP's required their inclusion in the proposals.

As can be seen, technical consultation, training and proficiency testing RFP's all required a description of the methods of laboratory selection and impact evaluation. Items three and four --"how proposed program will solve problem" and "purpose and nature of program" are conceptually similar and all three RFP's required one or the other. As can be seen in Table 4, CDC required that several evaluation-related items be described in one or two type(s) of proposals, exclusive of the other type(s).

The proficiency testing RFP was the most unique of the

Table 4
 CDC Proposal Specifications¹

Items Requiring Descriptions and Explanations	RFP Stipulation		
	PT	TC	TR
1. Problem*	-	+	+
2. Magnitude of problem	-	+	+
3. How proposed program will solve problem†	-	+	+
4. Purpose and nature of program*†	+	-	-
5. Specific program objectives	+	-	-
6. Anticipated problems with implementation	+	-	-
7. Method of laboratory selection*	+	+	+
8. Target population character- istics--personnel*	-	-	+
9. Target population character- istics--institution*	+	+	-
10. Public relations and participant recruitment*	+	-	+
11. Needs assessment*	-	+	+
12. Formative evaluation*	-	-	+
13. Performance indicators to be used*	+	-	+
14. Methods to assure quality of tests (PT)*	+	-	-
15. Grading criteria --scoring methods*	+	-	-
16. Impact evaluation*	+	+	+
17. Contents to be included in final report	-	+	-

¹Sources: Center for Disease Control Requests for Proposals, Nos. 200-79-0911(P), 200-79-0912(P), and 200-79-0913(P), March, 1979, Technical Proposal Instructions Section B2.

*Related to guidelines appearing in the checklist.

†These two items are conceptually similar.

three in terms of its proposal requirements, probably because proficiency testing is perceived as more an assessment activity than an educational treatment (Forney and Brooke, 1967). Shepard (1977) points out that there is some awkwardness in applying evaluation criteria to the "evaluation of an assessment, since assessment is both the object of an evaluation and an evaluation activity itself." The incongruity found among the three RFP's prompted further investigation and formal hypotheses were drafted.

Inferential Statistics

The first hypothesis to be tested is stated as follows: There is no difference in total checklist compliance scores between the three types of contract proposals, i.e., proficiency testing, technical consultation, and training. To test the hypothesis, a Kruskal-Wallis one-way analysis of variance by ranks was performed (Siegel, 1956, pp. 184-194). The result is as follows:

$$\underline{H} = .17, \underline{p} > .15$$

The null hypothesis cannot be rejected. The three proposal types do not differ significantly on overall technical adequacy.

The second hypothesis relates to the scores of the three groups of contract proposals on each individual checklist item. The following null hypothesis is to be tested: The three types of program proposals do not differ with regard to their compliance levels on individual checklist items. This hypothesis, as in the first hypothesis, was derived from the prior observation that CDC's RFP's

specified some unique and some common items. It was believed that the unique demands of the RFP's would be related to differential scores on individual items among the proposals. To test the second hypothesis, a Mantel-Haenszel Chi Square Procedure (Mantel and Haenszel, 1959) was used. This procedure derives a contingency table for each checklist item, then summarizes over each of the cases to give a summary chi-square. The major advantage of the test is its extension to orderable test factors at more than two levels. The summary statistic was again not significant; however, several significant associations of program type to individual checklist item scores were found. Pertinent items appear in Table 5, grouped according to the corresponding CDC RFP stipulations. Of the 13 results either significant or approaching significance, it is interesting to note that eight relate directly to CDC RFP requirements (see Tables 4 and 5). This suggests that proposals within a program class tend to comply with the unique RFP demands. If the RFP does not stipulate the item as a requirement, the other proposals tend not to address the item.

The third hypothesis is derived from the observations in Table 5. It is stated as follows: Differences among the three proposal types on individual checklist item compliance are not related to the requirements set forth inconsistently by the RFP's. To state this another way, the alternative hypothesis is that when a checklist item is consistently required or not required by all three types of RFP's, there will be fewer significant differences in compliance to the checklist items than when the checklist items are only required

Table 5

Associations between Type of Proposal and Checklist Compliance
as a Function of RFP Requirements

Items Required of One Type Only	p Value	Items Required of Two Types	p Value	Items Required of All Three Types	p Value	Items Not Required of Any Types	p Value
Description of Target Population (Target Population Characteristics --Personnel)	.01	Plan for Cooperation/ PR (Public Relations and Participant Recruitment)	.04	Impact Evaluation Plan	NS	Plan to Evaluate Long Term Effects	.03
Formative Evaluation Plan	.12	Assessment Plan --Manifest Needs (Needs Assessment)	.004	Method of Selection (Method of Laboratory Selection)	NS	Reactivity of Measurement	.02
Identification of Instruments (Grading Criteria--Scoring Methods)	.04	Assessment Plan --Perceived Needs (Needs Assessment)	.004	Defined Purpose (Purpose and Nature of Program)	NS	Statistical Tests	.03
Judgment Criteria (Grading Criteria)	.0004	Description of Target Setting (Target Popu- lation Characteristics --Institution)	NS			Unit of Analysis	.08
Reliability (Methods to Assure Quality of Tests --PT)	.004	Measurement Methods (Performance Indicators to be used)	NS			Data Management	.05
		Clear Need (Problem)	NS			Remaining 12 Items	NS

Note. Items in parentheses indicate the distinct wording used in Table 4 for the RFP requirements. Numbers indicate p values of the associations where $\leq .05$ is considered significant and $\leq .12$ is considered approaching significance. NS = not significant.

by some of the RFP's. The two-way chi square test was used to test the hypothesis (Bartz, 1976, pp. 297-303). The results appear in Table 6. The chi square was significant at the .05 level and allows the null hypothesis to be rejected. The differences noted among types of contract proposals appear to be related to the funding agent's requirements.

Table 6

Association between RFP Requirements and Differences
in Compliance to Checklist Items

	Number of Checklist Items <u>In-</u> consistently Required by the RFP's	Number of Checklist Items Consistently Omitted (or Required) by RFP's	Total
Significant Differences in Compliance	8	5	13
Nonsignificant Differences in Compliance	3	12 (3) ^a	18
Totals	11	20	31

Note. Chi square = 4.82, $p < .05$.

^a Number in parenthesis indicates items that were required of all three program types by the RFP's.

Since the total scores of individual contractors vary considerably, additional available data were explored for relationships which might point to causes for the variability and useful predictors

for the higher levels of technical quality. The last two columns of Table 2 show the amount of contract funds awarded and the contractors' prior experience with federally funded laboratory improvement programs.

The final two hypotheses investigate the relationship between checklist scores and amount of contract money awarded, and checklist scores and previous years of experience adjusted for the number of contracts held. (See last two columns of Table 2.) The null hypotheses are as follows:

1. There is no relationship between compliance scores, as measures of technical quality of proposals, and amount of contract money awarded.

2. There is no relationship between checklist compliance scores and amount of prior contract experience.

To test the hypotheses, the Spearman Rank-Difference Correlation was calculated (Bartz, 1976, pp. 200-295). For compliance vs. funding $r_s = .28$, $p > .05$. Therefore, the null hypothesis cannot be rejected. For the second relationship hypothesis, compliance vs. experience, $r_s = .07$, $p > .05$. Again the null hypothesis cannot be rejected. Technical adequacy of the proposals was not shown to be related to the amount of funds awarded or the contractor's prior experience.

Conclusions

It can be concluded that although no single proposal ostensibly complied with all thirty-one checklist guidelines, the

aggregate group of proposals optimally fulfilled thirty of the guidelines and partially met the one remaining guideline (estimation of validity). Every proposal rated high on at least a third of the guidelines. The overall technical quality of the proficiency test, technical consultation and training proposals was equal among the three groups. Though no group of proposals outshined the others in total checklist scores, there were significant differences (or approaching significance) among the three groups on thirteen individual items. The data indicated that differences were associated with the funding agent's stipulations as they appeared in the RFP's. Other variables were examined for their relation to the technical quality of the proposals. Neither the amount of contract funds awarded¹ nor the amount of the contractors' prior contract experience appear to be important predictors of technical adequacy.

The following discussion is divided into two subsections: (1) general interpretations of the study, and (2) interpretations of specific proposal deficiencies.

General Interpretations of the Study

In general, the data suggest that the funding agent has the greatest influence on the technical quality of the proposals by

¹To increase the validity of measuring the amount of contract funds awarded, appropriate adjustments should be made to account for the number of laboratories (or individuals) to be enrolled, distance and amount of traveling required, frequency and duration of contacts with participants, and so forth. This kind of data was either unavailable or incomplete in the portions of the proposals reviewed. This information was not necessarily omitted. It may simply have appeared in a section other than that which was requested (of the funding agency) for this review.

stipulating certain factors in the RFP's and presumably selecting contractors who meet those requirements. The fact that CDC set forth some specific directives relating to evaluation in their RFP's is a positive step in the right direction. According to Scriven (1976, p. 121), "The first great step toward accountability consisted of requiring that there be some evaluation of tax-funded or foundation-funded projects." The twenty-three proposals that were awarded funds have passed the first test and the subsequent success or failure of their programs will be in part, an evaluation of the funding agent itself (Scriven, 1976). Both parties, the funding agent and grantee, have a vested interest in program success. Therefore, the act of stipulating certain elements of program evaluation in the RFP's lends credibility to an otherwise biased situation.

This discussion is not complete without some reflection on plausible alternative explanations for the observation that certain types of proposals were more attentive to particular evaluation guidelines than the other(s). Other causes for significant differences among proposal types on particular checklist items are possible. The RFP requirements appear to be associated with the differences among proposal types, but the associations may be the result of some external cause(s). Since five of the thirteen significant differences were not directly related to RFP requirements, some additional independent variable warrants consideration. The significant differences in item quality could have been due to differences indigenous to the particular orientation of the type of program proposed,

e.g., proficiency testing is more assessment oriented and may naturally be strong on instrumentation items and weak on program context; technical consultation is more oriented to problem identification and amelioration and may be innately stronger on pinpointing clear need and manifest needs, while less adequate on uncovering perceived needs and participants' preferences. Training is more concerned with information dissemination to individuals and would consequently need a willing group of participants and some clear-cut ideas of what information should be distributed--it makes sense that the strength of training proposals would be in context items.

To conclude that the funding agency is in the best position to upgrade the technical quality of laboratory improvement programs is tenable in any case, since the criteria for awarding a contract or grant must be set and enforced by the funding agent. The funding agent undoubtedly has the best vantage point. The most convincing admonitions in guidelines and standards will go unnoticed or unheeded without some strong incentive to make the extra effort remunerative. The funding agent is solely in control of that primary incentive.

Interpretations of Specific Proposal Deficiencies

There is still a long way to go before confident decisions about laboratory improvement programs can be made based on valid evaluation data. The constituents of the context, structure and instrumentation categories of evaluation must be more carefully planned in the future. The analytical review in this chapter has

uncovered several weaknesses of proposed laboratory improvement program evaluation under each of the three categories.

Some of the most obvious flaws in terms of context were as follows:

1. Many proposals seemed to confuse problems with solutions, as discussed in Chapter 2. The lack of regulations, training or voluntary enrollment in proficiency testing was cited as a primary justification for program installation. This is not the problem, but just a different way of treating the problem.

2. Several contractors who proposed to develop technical consultation programs cited laboratories' general lack of awareness of limitations as the major cause of poor quality laboratory results. Yet no mention was made of the strategy to be used to teach laboratorians their limitations.

3. The training proposals in general had very thorough needs assessment plans. The only flaw seemed to be in articulating what would be done with all the information. For data to be valuable, there must be a plan to use them.

In general, the structural aspects of laboratory improvement program evaluations are weak. Specific deficiencies noted seem to be due to misunderstanding and lack of familiarity with the more abstract evaluation concepts and principles. Several contractors did not seem to understand the different types of evaluation called for by the training RFP. "Evaluation of student performance," (formative) "system for monitoring/improving training," (program monitoring) and "evaluation of the training on laboratory performance,"

(summative) were all blended together. Several contractors also seemed to have difficulty interpreting the RFP words "educational objectives."

The most striking inadequacies that fall under the instrumentation category relate to validity and reliability of evaluation measurements. For example, the popular systems for scoring Bacteriology PT results seemed to have no grounding in the reality of host-parasite relationships. The commonly used formula,

$$\frac{\text{number of appropriate responses}}{\text{number of correct organisms} + \text{number of errors}}$$

assumes that the danger of finding too many is equal to the danger of detecting too few organisms. In many cases, this just is not true. Far better it is to find a few Beta hemolytic streptococci and erroneously report a few Staphylococci than to be able to see all the normal flora, but overlook the Beta streptococci. The same logic applies to scoring of antimicrobial susceptibility testing as the:

$$\frac{\text{number of appropriate determinations} \times 100}{\text{number of agents tested}}$$

This formula will obfuscate a major error that could have serious implications for patients. For example, with this formula, if twelve drugs are tested and one--perhaps the drug of first choice--is erroneous, the laboratory still achieves greater than a 90 percent score. There is little regard for the effects of such an error on real patients.

Another validity problem was noted on several of the training

proposals which lacked any plan to measure on-the-job performance. Instead, written tests or self-reports would be relied on to yield data demonstrating the effectiveness of training in improving laboratory performance. In this regard, laboratory improvement programs and continuing education programs for health professionals discussed in Chapter 2 are very much alike. The inferential leap from paper and pencil measures to actual job performance is alluring despite its dubious validity.

Reliability is the most noticeable instrumentation weakness in onsite technical consultation programs which intend to make observations about laboratorians' behaviors and conditions of the work environment. Interrater reliability and internal consistency were not considered in any of the technical consultation proposals, although one contractor appeared to at least have the potential to determine interrater reliability.

This conclusion section has reviewed the major findings of the entire meta analysis reported in this chapter. The general study findings were interpreted first followed by some suggested alternative explanations. This was followed by interpretations of specific proposal deficiencies. The technical quality of laboratory improvement program evaluation must advance to the level of currently accepted evaluation principles before valid inferences about program impact can be made. The funding agency is in the most advantageous position and has the authority to accelerate the process.

In summary, this chapter has presented a thirty-one item checklist along with detailed descriptions and rating scales

intended to facilitate further use of the items by program evaluators, project directors, granting authorities, and proposal writers. The checklist was field tested on twenty-three contract proposals of the most current federally funded laboratory improvement programs. This use of guidelines was referred to as formative meta analysis because it examined the potential technical quality of evaluations planned by laboratory improvement programs. The following chapter will present a summative meta analysis since it sums up the overall merit of completed laboratory improvement program evaluations (Stufflebeam, 1978, p. 23).

The results from this chapter's meta analysis can be used to provide direction and guidance to those involved in future program development at the federal and state level. The purpose of this analytical review of contract proposals was to explore the current state of evaluation thinking that will be applied to laboratory improvement programs; and to provoke careful reflection and useful creativity among those who will continue to contribute their welcome energies to the progress of laboratory medicine and health care delivery.

Chapter 4

A SUMMATIVE META ANALYSIS OF EVALUATIONS COMPLETED BY FEDERALLY FUNDED LABORATORY IMPROVEMENT PROGRAMS

This chapter is the logical sequel to the formative meta analysis presented in Chapter 3. Whereas a formative meta evaluation assesses the extent to which certain guidelines are met in the planning stages of evaluation, summative meta evaluation assesses whether standards were adhered to in a completed evaluation. Stufflebeam (1978, p. 23) describes summative meta evaluation as a means to

. . . hold evaluators accountable by publicly reporting on the extent that their evaluation reports meet standards of good evaluation practice . . . [and] help the audiences of primary evaluations determine how seriously they should take the . . . reported conclusions and recommendations.

This chapter will discuss indepth two completed laboratory improvement program evaluations to trace the thread of the technical dimension of evaluation as it winds through their final reports. This analytical review will use real examples of completed evaluations and expound on the more abstract technical evaluation concepts as they were described in the guidelines in Chapter 3. The concepts examined in this review include needs assessment, evaluation design, methods of selection and assignment, internal and external validity, statistical tests, and instrument validity and reliability. Rather than discuss these concepts separately, they are integrated into the narration about the two completed evaluations

where appropriate, as they relate to activities, results, and interpretations, in that order. It is the intention of this chapter to make the abstract technical evaluation concepts that were examined cursorily in Chapter 3 more intelligible and thus more usable. The idealistic mien of the checklist will be seen from a more utilitarian perspective, while preserving and reinforcing the universal value of adherence to evaluation guidelines like those in Chapter 3.

In addition to reviewing the evaluation concepts described previously under checklist items in Chapter 3, this chapter will cover some of the standards proposed by the Evaluation Research Society (ERS, 1980) under the categories of data analysis and interpretation, and communication and disclosure. This will broaden the evaluation perspective and assist laboratory improvement program evaluators who are charged with the responsibility to report on their program's effectiveness, so that the funding agency can make decisions whether to expand, discontinue or reexamine their approach. The relevant factors under the ERS categories of data analysis and interpretation are as follows:

When quantitative comparisons are made (e.g., X is greater than Y) tests of statistical significance should be applied and interpretations should be stated with some indication of confidence.

Cause-and-effect interpretations should be bolstered not only by reference to the design but also by recognition and elimination of plausible rival explanations.

Findings should be reported in a manner that distinguishes among objective findings, opinions, judgments, and speculation. (ERS, 1980, p. 18)

The relevant factors under communication and disclosure include:

Limitations caused by constraints on time, resources, data availability, etc. should be stated.

Assumptions should be explicitly acknowledged.

Findings should be presented clearly, completely, and fairly. (ERS, 1980, pp. 19-20)

As was the case for the evaluation concepts taken from the checklist, these standards also will not be discussed independently, but the essence of the standards will be incorporated into the narration as appropriate to the contractor's reports of their completed evaluations.

Valid inference, which is the main thrust of the technical domain of evaluation, is not the only concern for decision making purposes. Nontechnical aspects such as cost benefit, responsiveness, and probity, must be included in the total picture (Stufflebeam, 1978). These parameters were not addressed in this review. The focus here was on the basic minimum--the technical aspects of evaluations.

The combination of this chapter and Chapter 3 represents the entire spectrum of laboratory improvement program evaluation and the current state of its technical quality. The previous chapter summarized the common shortcomings of evaluation plans and this chapter will illustrate the common pitfalls in evaluation implementation and in reporting program effectiveness. This chapter is divided into sections bearing the subheadings (1) selection of two example programs, (2) evaluation example I, (3) evaluation example II, and (4) summary.

Selection of Two Example Programs

The two examples of completed laboratory improvement program evaluations were chosen to represent the two extremes or anchors on the continuum scale of interpretable evaluation designs. The continuum scale is based on degree of experimental rigor. On the one extreme lies evaluation designs allowing valid inferences about cause and effect, e.g., training caused improved laboratory performance. Programs located at the other extreme are limited by their nonrigorous evaluation designs to extremely cautious conclusions about program effects; causation cannot be inferred from the data available.

It cannot be said that designs at the one end of the continuum are singularly perfect and designs at the opposite end entirely corrupt. Each may have particular strengths and weaknesses. In a discussion of organizational research, Homans (1962) said,

People who write about methodology often forget that it is a matter of strategy, not of morals. There are neither good nor bad methods, but only methods that are more or less effective under particular circumstances in reaching objectives on the way to a distant goal. (p. 257)

The two laboratory improvement programs reviewed showed many fundamental similarities. The type of treatment studied was technical consultation. The format of the technical consultation was the same. The two programs were located in adjacent states with similar populations and ratios of urban to rural communities. There was some difference in the target populations. One study was directed at physician office laboratories, whereas the other study included laboratories of many types and sizes.

Evaluation Example I

Target Population and Goals

This program was offered to physician office laboratories, which were recently mandated--by their state legislatures--to establish quality control systems and participate in an approved proficiency testing program ("An Analysis of a Laboratory," 1979a).

The major goals of the program were:

1. To uncover discrepancies between the performance observed in these laboratories and the standards imposed by state regulations.
2. To correct any deficiencies and improve laboratory proficiency test performance through onsite technical consultation.

The study group consisted of 109 physician office laboratories. The design included ten nonphysician office laboratories which the contractor designated the control group. The total target population included 135 physician office laboratories; of these, 109 laboratories volunteered to participate. The ten laboratories in the control group were selected by the contractor. Nine small hospital laboratories and one independent laboratory comprised this group. The basis for their selection was prior compliance to regulations. The ten control group laboratories were visited by consultants during the same time the study group laboratories received their first treatment visits. The ten control group laboratories were only evaluated; no consultation was given.

Treatment and Needs Assessment Activities

A rating instrument was developed for use during the onsite interviews which were to be conducted by the technical consultant. The interviews served as the springboard for consultative advice whenever a deficiency was noted. A follow-up letter reemphasized the deficiencies. The rating instrument did not measure technical performance. It focused on cognitive issues. The consultant administered an onsite proficiency test concurrently with the interview. No checklist or other instrument was used to record behavioral observations. The formats of the consultation were varied depending on the discrepancies observed. The consultant spent much of the time urging laboratories to comply with regulatory standards, adhere to protocols, and use information contained in package inserts. In some instances, the consultant contacted salesmen on behalf of the laboratories visited, requesting them to replace expired reagents or provide more up-to-date products.

Evaluation Design

The consultants revisited 57 of the 109 laboratories between two and five months after the first visits. These 57 laboratories were selected from a total 87 laboratories that the contractor felt were significantly deficient on the first visit. The method of selection was not disclosed. The purpose of the second visits was to evaluate changes brought about by the program of technical consultation ("An Analysis of a Laboratory, 1979a, p. 11). The laboratories' follow-up responses to the interview items were compared

to their responses from the first visit.

There are several methodological problems with the evaluation design up to this point. To begin with, external validity could be threatened by the fact that only volunteers participated in the study; generalization to all physician's office laboratories would thus be impossible. However, the contractors were able to recruit 109 out of 135 labs, or 81 percent of the total available. Since this is a relatively high percentage, it adds considerable credibility to the potential external validity of the design (Rosenthal and Rosnow, 1975). External validity is usually only an issue of secondary importance compared to internal validity (Campbell, 1969, pp. 165-185). According to Campbell and Stanley (1963, p. 5), internal validity is the "sine qua non . . . without which any experiment is uninterpretable."

Thus, the internal validity of this contractor's evaluation design was a more important issue. It originally appeared that the contractor planned to compare the performance of two groups--the group that received consultation and the control group that did not. Yet the contractor made no further mention of the control group laboratories in any sections of the report following the study design section. No data were given on the performance of the ten control laboratories. This was a most curious hiatus between the apparent evaluation plan and the actual practice. The reasoning behind the inclusion of the ten control laboratories in this study was a mystery. It was later solved in a different report by the same contractor regarding another concurrent study. The control

group was used to see whether physician office laboratories performed as well as small hospital laboratories, which had been under state regulations longer ("An Analysis of Idaho," 1979b).

Since a between group comparison was not done for this study, the design essentially boiled down to a one group before and after (pretest-posttest) study (Campbell and Stanley, 1963, p. 7). As such, there are several variables jeopardizing any internal validity as mentioned in Chapter 3. The fact that 57 of the worst laboratories were selected for reevaluation, out of the 109 laboratories initially evaluated, makes the results vulnerable to the insidious effects of statistical regression toward the mean. It is possible that the better laboratories observed on the first visit would do worse on the second visit, just because of the inevitable imperfections of subjective interviews. On the other hand, the poorer performing laboratories might do better on the second visit, for the same reasons. If both groups had received the post evaluation, the total effect would cancel out the worst group's improvement.

The one group before and after evaluation design also does not prevent the threats of history and maturation (Campbell and Stanley, 1963, pp. 7-9). There could be other programs such as inservice education or training through a private enterprise--going on concurrently with the study--that could cause improved performance. Laboratories also may improve or get worse because the passage of time alone influences behavior.

The final problem with this evaluation design is the strong possibility for observer bias (Borg and Gall, 1979, pp. 159-162 and

523; Gronlund, 1976, p. 442) and experimenter expectancy (Fromkin and Streufert, 1976, pp. 439-441) to cause spurious results and conclusions. Observer bias was discussed in Chapter 2. Experimenter expectancy is related to observer bias but refers more to the effect of the observer's subtle cues on the behavior of the participant. Observer bias relates to errors in observer's judgment whereas experimenter expectancy has to do with errors in the participant's responses.

Results and Data Presentation

The results were tabled and excerpts can be seen in Table 7 and Table 8. The use of percentages instead of a simple count of the number of laboratories is potentially misleading. In Table 7, for example, what is listed as a 35 percent improvement under the Quality Control Adequate item actually means that about 17 laboratories complied with the item on the first visit and 26 laboratories complied with the item on the second visit; nine laboratories changed from unacceptable to acceptable. The 35 percent improvement could be misinterpreted as meaning that, in general, laboratories performed 35 percent better on the second visit. A more accurate column heading would have been Percent of Total Labs That Changed from Unacceptable to Acceptable. Using the number of laboratories instead of the percentage would have been even more straightforwardly interpreted, but a 35 percent improvement does sound more impressive than nine laboratories improved. The same is true for Table 8. A 33 percent improvement actually means only one laboratory changed

Table 7
Changes in Laboratory Performance in Chemistry
Post Technical Consultation

Chemistry Questionnaire Item	<u>First Visit</u> Percent Acceptable (N - Variable)*	<u>Second Visit</u> Percent Acceptable (N - Variable)*	Percent Improvement
Quality Control Adequate	65	100	35
Quality Control Results Recorded	74	100	26
Calibration of Spectrophotometer and Colorimeter Daily or as Used	56	87	21
Calibration Checks Recorded	50	68	18
Logbook for Preventive Maintenance	45	62	17
Logbook for Preventive Maintenance Up-to-Date	47	74	27
Procedure Manual of Inserts Available	96	100	4
		Average Improvement 25 percent	

*N - varied between 22-27.

Source: Excerpted from "An Analysis of a Laboratory Improvement Program for Idaho Private Physician Office Laboratories through Technical Consultation," Final Report of Contract #200-77-0742 to the Center for Disease Control Bureau of Laboratories, Atlanta, Georgia, 1970a, p. 29.

Table 8

Changes in Laboratory Performance in Immunology
Post Technical Consultation

Immunology Questionnaire Item	<u>First Visit</u> Percent Acceptable (N = 3)	<u>Second Visit</u> Percent Acceptable (N = 3)	Percent Improvement
Quality Control Used Both Positive and Negative	67	100	33
Are quality Control Results Recorded	0	100	100
Procedure or Inserts Available	100	100	0
Copy of Patient Results Kept in Laboratory	67	67	0
		Average Improvement	33 percent

Source: Excerpted from "An Analysis of a Laboratory Improvement Program for Idaho Private Physician Office Laboratories through Technical Consultation," Final Report of Contract #200-77-0742 to the Center for Disease Control, Bureau of Laboratories, Atlanta, Georgia, 1979a, p. 42.

from unacceptable to acceptable.

The attempt to summarize the data by the calculation of average improvement is also misleading. The contractor simply averaged all the percentages appearing in column four. For Table 7 the number of laboratories in the denominator is different for some of the items; averaging the percentages rather than the raw numbers gives an inappropriate weighting to the data.

If simple counts of laboratories had been used, McNemar's test of changes (Bartz, 1976, pp. 313-314) could have been done for each item to determine the significance of the association between performance and a technical consultation visit. For this to be done it would be necessary to know the number of laboratories (if any) that changed in the opposite direction, i.e., from acceptable to unacceptable. The contractor's report did not provide this information. It would also have been useful to know whether the total number of deficiencies per laboratory decreased from one visit to the next. A Sign Test (Bartz, 1976, pp. 314-316) would be used to test significance in this case.

Interpretations

The contractor's report states ("An Analysis of a Laboratory," 1979a, p. 60), "The tables show that a significant improvement occurred in all areas of laboratory performance in the fifty-seven laboratories receiving technical consultation." Actually, of the 88 items listed, 11 showed no changes in performance when improvement was possible. Also, the data were not subjected to any statistical

tests of significance. There were no controls to rule out alternative explanations for improved performance. The certainty of the contractor's statement is overextended.

The contractor also provided some data on changes in proficiency test scores ("An Analysis of a Laboratory," 1979a, p. 58). Results can be seen in Table 9. From the data, the contractor concluded that there was a trend toward improvement (p. 59). The order of magnitude of the difference in percentages is relatively small, but in order to accurately interpret this, additional data would be necessary. They could not be found in the report. First, the number of laboratories performing in each discipline was not given. It appeared that the error rates were not just representative of the 57 laboratories who were given consultation and revisited; they encompassed all 135 physician office laboratories in the entire state ("An Analysis of Idaho," 1979b, pp. 3-4; "An Analysis of a Laboratory," 1979a, p. 4). Applying the data to infer the performance changes of the subgroup previously designated as the worst laboratories is inappropriate.

If the proficiency test results were valid indicators of the treated laboratories, the next necessary piece of information is the standard deviation of the error rates. This would allow calculation of a correlated t -test (Bartz, 1976, pp. 259-263) if the basic assumptions underlying the t -test were met (p. 253). If the assumptions were not met, for example, if the standard deviation of the 1977 error rate was more than twice as large as the standard deviation of the 1978 error rate, then the nonparametric Wilcoxon

Table 9
Changes in Proficiency Test Error
Rates Over One Year

Discipline	Error Rate		
	1977	1978	Difference
Urinalysis	8.6%	7.2%	-1.4%
Hematology	10.8%	8.8%	-2.0%
Chemistry	11.2%	10.2%	-1.0%
Bacteriology	17.8%	18.8%	+1.0%

Source: "An Analysis of a Laboratory Improvement Program for Idaho Private Physician Office Laboratories through Technical Consultation," Final Report of Contract #200-77-0742 to the Center for Disease Control, Bureau of Laboratories, Atlanta, Georgia, 1979a, p. 29.

Matched-Pairs Signed Ranks Test would be more appropriate (Bartz, 1976, pp. 253 and 316-319).

The inferences from any of these statistical tests suggested are still susceptible to alternative explanations due to a weak evaluation design. To demonstrate that technical consultation causes either fewer deficiencies or lower proficiency test error rates, a program must control for the many other variables that may explain improvements in performance. There are studies that suggest proficiency test participation over time is by itself related to improved performance (Finkel and Miller, 1973; La Motte, 1977; Peddecord, 1978, p. 12). Improvement in proficiency test performance may occur because of the educational effects of the feedback

given to participant laboratories, or because of test-wiseness, the term Borg and Gall (1979, p. 523) used to describe a subject's experience acquired with repeated testing. Test-wiseness plagues internal validity. The contribution technical consultation alone made to improved laboratory performance is indeterminate, regardless of any manipulation, statistical or otherwise, of the data from this study.

Insights

The contractor did provide a thorough description of the most commonly observed laboratory deficiencies. This information is useful for others who may need to develop or adapt measurement instruments for rating laboratories. Also well described were the circumstances under which laboratorians graciously accepted advice and those instances where gentle pressure met considerable resistance. Their experience with persuasive techniques would be valuable to programs and individuals aspiring to improve laboratories.

Several recommendations were provided for consideration by future technical consultation programs:

1. Visits for evaluation should be separate from consultation visits to allow more instructional time.
2. Written correspondence intended to follow-up on deficiencies cited and suggest improvement strategies should be brief and plainly worded. Indepth discussions about complex topics such as quality assurance are better handled face to face.
3. To prevent inconvenience to laboratory personnel, visits should be planned during off or slack hours.

Evaluation Example IITarget Population and Goals

As in the previous example, this study sought to assess and improve laboratory performance through onsite technical consultation ("Final Report," 1979). The target population consisted of the 70 laboratories that were all formally accredited by an independent (government or private) agency and participated in proficiency testing programs in Chemistry, Bacteriology or Parasitology. These three disciplines showed many laboratories performing at unacceptable proficiency test levels. Like the previous study, the objective of the technical consultation was to explore possible causes and facilitate an upgrading process. Unlike the previous study, where the target population was a homogeneous group, this program encountered a great deal of diversity in the institutional settings and staffing structure of the laboratories enrolled.

The study sample included 39 hospitals of various sizes, out of the 41 hospitals existing in the state; 12 physician clinics each serving more than five physicians, out of 16 existing clinics; and three interstate licensed independent laboratories of the 13 eligible independent labs. These 54 laboratories were selected using a table of random numbers. Stratified sampling was performed not by size or type of laboratory but by an economic priority; the purpose was to include as many laboratories who performed testing in all three disciplines as possible. This would allow a minimum number of laboratory visits with maximum data yield. Due to this

strategy, independent laboratories were underrepresented and hospital laboratories were overrepresented.

Treatment and Needs Assessment Activities

Two different forms of technical consultation were tested in this study. One required initial onsite proficiency testing while the technical consultant observed and rated procedural performance according to 20-item checklists condensed from the College of American Pathologists (1974), Center for Disease Control, and Medicare (U.S. Department of Health, 1978) survey instruments. Performance was judged as compliant or noncompliant according to present criteria on the checklists. After the needs assessment, the consultant provided immediate feedback to the laboratorians on the accuracy of their proficiency test results and any deficiencies observed during the testing. Recommendations were given, a copy of the checklist standards was distributed along with explanations, and demonstrations were presented when deemed necessary.

The alternate method of technical consultation was less intrusive. The checklists and proficiency test specimens were delivered to the laboratories along with verbal instructions: process the specimens, read and complete the checklists by rating your (the laboratorian's) perceptions of how you comply with the items, and include any explanatory or critical remarks. This was done primarily to insure that the checklists would be read thoroughly and related to the laboratory's routine performance. It was hoped that this self-assessment approach would improve performance. The

consultant did not interact with these laboratorians. A self-addressed stamped envelope was given to these laboratories along with the request that they return their results to the consultant. No feedback about results was given.

Evaluation Design

Stratified-random selected laboratories were assigned to three groups using a table of random numbers. The three groups included one no-treatment control group and two variations of the technical consultation intervention. Laboratories were recruited by telephone and all laboratories contacted agreed to participate. Thus the external validity was maximized by random sampling and 100 percent cooperation. To determine the impact of technical consultation, laboratories in the two treatment groups were reevaluated twenty weeks after the first visits. The consultant again rated them on compliance to checklist standards and proficiency test accuracy. The control group laboratories were evaluated in the same way. The performance of the treatment groups after technical consultation was compared to the performance of the control laboratories who had not received technical consultation.

The design was a posttest-only control group design, which qualifies as a true experimental design according to Campbell and Stanley (1963, pp. 25-27). Campbell (1969) extolls this design although many educational researchers often regard it disdainfully because of its lack of baseline, pretest data. This design is preferred over the more traditional before and after control group

design because of its control for the testing-treatment interactions that threaten external validity (Campbell and Stanley, 1963, p. 25). The inferential statistics for the pretest-posttest control group design are more powerful than those for the posttest-only control group design, however, and to overcome lack of power, Campbell and Stanley recommend blocking on antecedent variables or using them as covariates (p. 26). In this study, laboratory weekly or monthly test volume was used as a covariate. An advantage of the use of a covariate in this study was the contractor's ability to examine interactions of consultation and test volume. Covariance can address the question, "What is the effect of workload on the level of laboratory performance and do laboratories improve after consultation differentially according to their workloads?" ("Final Report," 1979, p. 23).

Results and Data Presentation

The results of this study, based on an analysis of variance, showed no significant differences between treatment and control laboratories on either compliance to standards or onsite proficiency testing scores. The contractor interpreted this as a Type II error or failure to find a difference that truly exists, rather than the ineffectiveness of technical consultation in improving laboratory performance (p. 7). They offered no justification for that conclusion.

The contractor carried out additional statistical manipulations to study the relationships among individual checklist items.

Relationships were also explored between checklist compliance and proficiency test scores (pp. 8-25). Their data from the consultant's ratings on discrete chemistry tests (glucose, bilirubin, sodium, BUN) were correlated and then factor analyzed. Intercorrelations among the different test procedures showed four factors that were consistent across the types of chemistry tests: parts of the quality control system that relate to random error, such as within run reproducibility checks and establishment of normal ranges; parts of the quality control system that relate to systematic error, such as checks for between run reproducibility and verification of results prior to reporting patient values; and preventive maintenance. In Bacteriology, the individual checklist items loaded onto twelve factors; and in Parasitology, seven factors were generated. The twelve factors in Bacteriology included following written and referenced protocols, documenting judgment criteria, ability to recover anaerobes, preventing sources of error by quality controlling biochemical tests, and using a variety of media to maximize recovery of an organism from a specimen. The important factors in Parasitology included the use of basic procedures such as concentrations of stools and interpretive aids such as a Parasitology Atlas, documentation of all observations, and proper microscopy (pp. 9-16).

Factor scores were then derived for each laboratory. Multiple regression was undertaken to determine the set of factors that best predicted the onsite proficiency test outcomes. Mailed proficiency test scores were included in the regression procedures

(pp. 19-22). None of the factors from the Chemistry ratings predicted scores of mailed proficiency tests, although several factors were useful in predicting onsite proficiency test scores. The scoring system for the onsite proficiency tests dichotomized results as acceptable or unacceptable. The factor scores were most useful in predicting onsite PT scores for specimens in the normal as opposed to abnormal range. Quality control of systematic and of random error appeared to be the most consistent predictors across the different analytes (Glucose, BUN, etc.) and level of analytes (normal vs. abnormal).

In Bacteriology, three factors predicted mail distributed PT scores and five factors predicted onsite PT scores. Two of the predictors were the same for both types of PT: use of standard protocol and documentation of judgment criteria.

For Parasitology, factor scores could not be derived because intercorrelations among factors were too high. However, the contractor performed a stepwise multiple regression on the individual checklist items vs. the PT scores and found no subset of best predictors (p. 19).

The contractor also computed intercorrelations among the onsite PT scores and mail distributed PT scores of the laboratories in this study. Several small but significant correlations were found among the scores. An interesting finding was that scores from onsite proficiency test specimens in the normal range did not correlate very highly with abnormal range PT scores. This suggested to the contractor that perhaps:

. . . different skills and method idiosyncracies are involved in the performance of tests [on specimens] at normal and abnormal levels . . . complete evaluation of a laboratory ought to include . . . different levels of the test agent. ("Final Report, 1979, p. 23)

Finally, the contractor examined the relationships between laboratories' test volumes in a particular discipline and their performance on the onsite PT specimens, and test volume vs. compliance to checklist items. The correlations between test volume and compliance to checklist items was high in all cases indicating that the more tests performed in a discipline, the more likely the laboratory conformed to standards in its procedures. The Pearson r correlations between total volume and onsite proficiency scores were less impressive except for Bacteriology and Parasitology where the correlations were .54 and .41, respectively; these are significant at the .01 and .05 respective levels. The contractor did not provide correlation coefficients for the relationship between test volume and mailed proficiency test scores. This was disappointing because it could have corroborated or refuted findings published by other researchers such as Finkel and Miller (1973) and Peddecord (1978).

Peddecord (1978, pp. 65-74, 83-84) reported that workload and laboratory size were not related to mailed proficiency test performance in the areas of Chemistry, quantitative Hematology, Blood Bank, Immunology or Syphylis Serology. In his study of 40 military hospital and outpatient laboratories, Peddecord did uncover significant associations between size and mail distributed proficiency test results for qualitative Hematology, Bacteriology and Parasitology. He attributed the failure of size and other variables to

predict Chemistry PT performance to the widely disseminated standardized technology which enables more uniform levels of service in most laboratories regardless of size. The contractor in this example offered a somewhat different but related explanation for the lack of correlation between Chemistry onsite PT scores and checklist compliance, and between Chemistry mailed proficiency test scores and onsite PT scores. They suggest ("Final Report," 1979, p. 23) that replicate testing of proficiency specimens is commonly performed by laboratories to generate means closer to the true value. This uncustomary replication, of dubious virtue under any circumstances, was hardly possible under scrutiny of a representative from a regulatory agency (p. 34). On the other hand, the contractor observed that "bacteriology requires a high degree of pattern recognition and judgment skill . . . replications [are more] limited by available resources, i.e., personnel and materials" (p. 34). Thus the opportunities for multiple repeats in Bacteriology testing are rare.

Interpretations

Unlike the final report discussed in Example I, which ended on a positive self-aggrandizing note, this final report took a more self-critical and disparaging position. They discussed in detail their interpretations and possible alternative explanations. This was followed by a thorough disclosure of all the methodological problems and errors that occurred during their program implementation and evaluation phases. Several evaluation authors stress the

importance of being candid, honest, and even critical, if appropriate, when reporting evaluation results (Anderson and Ball, 1978, p. 151; Evaluation Research Society, 1980; Stufflebeam, 1978). In a rebuttal article aimed at breaking down program directors' resistance to experimental rigor in evaluations, Boruch (1976) said it requires intestinal fortitude on the part of program directors "to not only evaluate their intentions rigorously but to admit that their program is not especially effective" (p. 180).

The report in this example called attention to their lack of sufficient numbers of laboratories to attend to statistical power considerations. The contractor cited one important problem relating to the internal validity of their evaluation. They reported that the technical consultant was also the evaluator (as in Example I) and she obviously was not blind to experimental conditions, nor impartial to the success or failure of the treatment. They described a gradual drift toward greater severity in the observer's ratings. "The technical consultant noticed her ability to detect deficiencies improved with time" (p. 35). A quirk in the implementation of their evaluation design resulted in the evaluation of all the control group laboratories several months before either of the two treatment groups were evaluated. The treatment groups' performance was therefore downgraded on the criterion measures. This would exacerbate any detection of a significant difference in the desired direction.

An instrumentation problem they reported was an apparent invalidity of the chemistry checklists since the factor scores did not predict mail distributed PT scores. Other investigators

(Black et al., 1976; La Motte et al., 1977; McCormick, et al., 1978) have found discrepancies between mail distributed PT performance and blind PT performance. In light of this, the contractor suggested that other intervening variables be anticipated and measured or controlled in the future.

Another instrumentation problem cited was the insensitivity of the chemistry PT scoring system. The preassayed serum they used proved not to be amenable to the usual statistical analyses applied to interval data. The measures of inter-laboratory variation, as reported by the well known manufacturer of the serum product, were later discovered to be based on historical trends and the manufacturer's intuition rather than empirical observations (p. 35). This forced the consultant to score the results on a nominal scale. The power of the inferential statistics was thus decreased to an even greater degree. Reliability of the simulated specimens used for Bacteriology and Parasitology was also cited as a problem.

Other sources of error they reported in their assessment instruments were as follows: no field trials of the checklists prior to their use in the study, a lack of sensitivity of the nominal (yes/no) checklist ratings, oversimplification implicit in the checklist items, no prior verification of the PT specimens for content validity, and lack of standardized objective scoring criteria.

The final report went on to delineate the major threats to internal and external validity according to Cook and Campbell's (1976) comprehensive work. The contractor called attention to the participants' anxiety and apprehension during the onsite proficiency

test. Such stress, they noted, may have caused less than optimum performance or intentionally poor performance if the participants resented the intrusion (p. 40). In all, 13 different validity problems were discussed. There seemed to be a genuine concern for careful interpretation of the study's results.

The contractor attributed many of the problems and inadequacies to time constraints imposed by contract deadlines, difficulties with goal setting (Rossi, et al., 1979, pp. 58-60), and a lack of developmental perspective (p. 39). Hayman and Napier (1975, pp. 74-79) uncovered many of the same problems in outcome-oriented evaluations in the public schools. The end result, they said, is that good programs which may have very favorable but latent characteristics, are discontinued.

Insights

The contractor in this example recommended that future laboratory improvement programs recognize the role of the administrators, pathologist-directors, clinicians, accreditation agencies, and the health care consumer in affecting laboratory performance. The demands of these diverse groups all converge on the clinical laboratory work force. Laboratory improvement efforts must operate in harmony with all influential groups to be maximally effective. Conflicting goals alone can cause performance problems in laboratories (Krieg, et al., 1978, pp. 131-151). The final report from this study recommended that all of these groups either be involved in or solicited for endorsement of clinical laboratory improvement programs.

In the following statement, the contractor seemed to echo the sentiments of Robert Mager and Peter Pipe (1970) regarding work performance.

Reinforcement for proper performance and commensurate penalties (for negligence) are neither adequate nor appropriate in clinical laboratories. The contingents for adhering to established standards are too subtle and inconsistent to stand alone as primary motivators. ("Final Report," 1979, p. 39)

Summary

These two examples were discussed to contrast the evaluation methodologies of the otherwise similar laboratory improvement programs. The different strengths and weaknesses of these and other contract reports provided the major impetus for the development of the checklist guidelines in Chapter 3. The object of the detailed discussion of the two programs was to elucidate the gist of those guidelines requiring considerable understanding and personal judgment, e.g., evaluation design, internal and external validity, method of participant selection, and validity and reliability of measurement instruments. This review also intended to convey the relative importance of the guidelines in Chapter 3.

The basic premise underlying this chapter and Chapter 3 is that poorly conceived evaluation designs and measurement instruments spawn erroneous conclusions. Under these circumstances, decisions are either stalled or misguided; judgment is impaired. When the results of such evaluations are reported to be positive, as in the first example presented in this chapter, there is a strong temptation to believe them.

The agency wants favorable feedback about its action, the project wants the agency to think well of it . . . so the situation is one of highly favorable evaluation. Against this formidable alliance, the search for truth is a little short of soldiers. (Scriven, 1976, p. 122)

Unless evaluation guidelines are established, explicit, and adhered to by program staff, funding agents cannot expect reliable, usable results; laboratorians as program recipients and their clinician/patient clientele may not get the quality, cost effective health care for which they've invested so much time and tax money.

Given that laboratory improvement programs have operated and propose to operate without a credible evaluation plan, it appears expedient for the funding agency to set more specific standards and designate prerequisites for future contractors and grantees.

Chapter 5

SUMMARY AND DISCUSSION

The efficacy of most past and projected laboratory improvement programs is tenuous at best. The justification for the very existence of federally funded laboratory improvement programs is equally vulnerable. Amidst a myriad of programs is a dearth of sound evaluation. The words of Freeman and Sherwood (1969, p. 74) are appropos. "For the most part, the evaluation requirement has remained a formality; granting agencies have tended to overlook it in their frenzy to implement programs intuitively believed worthwhile."

This chapter will summarize the preceding chapters and discuss the implications of this thesis under the headings of (1) review of goals, purposes and procedures, (2) technical quality in laboratory improvement program evaluations, (3) a need for valid data in laboratory improvement policy--the role of the funding agency, and (4) implications.

Review of Goals, Purposes and Procedures

When careful evaluations are carried out, their results are often consigned to oblivion. In a review of national health programs, the methodological quality of the evaluation design was shown to have little influence on the utilization of the data for decision

making (Alkin, Daillak, and White, 1979, pp. 21-23). The fate of valid laboratory improvement program evaluation is jeopardized by the same lack of concern or lack of knowledge. This research has attempted to intervene by:

1. reviewing the literature in evaluation theory to introduce basic evaluation concepts and principles;

2. reviewing the literature in continuing health professional education and clinical laboratory evaluation and improvement to determine the state-of-the-art of evaluation in programs aimed at upgrading the quality of health care through the development of human resources;

3. developing a checklist of evaluation guidelines that can be used by program directors in planning and implementing evaluations, and by funding agents in awarding funds. To aid potential users, checklist items have been described in detail including explicit rating scales;

4. field testing the utility of the checklist on 23 proposed laboratory improvement program proposals and assessing their technical quality;

5. evaluating two completed laboratory improvement program evaluations to explicate the correct interpretation and relative importance of (a) the more abstract evaluation concepts presented in the checklist, and (b) evaluation factors to consider when reporting on completed evaluations; and

6. exposing the pitfalls in the implementation and reporting phases of program evaluation.

Technical Quality in Laboratory Improvement
Program Evaluation

The findings of this research are as follows:

1. All three types of laboratory improvement proposals (proficiency testing, technical consultation and training) lack adequate evaluation designs with internal and external validity; they lack any plans to pilot test their programs, validate their evaluation instruments, monitor their programs' implementation, measure their programs' long term effects, or commission an unbiased evaluator. Few described and justified the method by which laboratories would be selected and assigned to receive the program, or the statistical tests which would be used in drawing inferences as to program effectiveness. Other items found lacking were specific to a particular type of program proposal.

The mean total score on the checklist ratings was 32.5 out of 63 possible points, with a standard deviation of 6.6. Scores ranged from 21 to 46.5.

2. The overall technical adequacy of the evaluation methodologies proposed by three different types of laboratory improvement programs (proficiency testing, technical consultation, and training) does not differ significantly.

3. The three types of laboratory improvement proposals differ significantly on the adequacy of certain aspects of the evaluations planned. The differences are related to the requirements set forth in the funding agents' RFP's. Five of the thirteen significant items were specifically required by the funding agency

for one program type, but not the other two; three items were required for two program types, but not the third.

4. The amount of contract money awarded is not related to the technical quality of the evaluation proposed.

5. The amount of the contractors' prior contract experience is not related to the technical quality of the evaluations proposed.

6. Past program evaluations differed with regard to the rigor of their designs and validity of their interpretations. The most extreme programs, at the far ends of the continuum of experimental rigor, had conflicting conclusions--one strongly in favor of the program; the other doubtful. The program with the rigorous design was more candid in disclosing weaknesses in their evaluation.

A Need for Valid Data in Laboratory Improvement Policy
--The Role of the Funding Agency

The interpretations of the findings can be summarized as follows:

1. Laboratory improvement program evaluation needs to improve in technical quality before any sensible conclusions can be made or rational policy enacted.

2. Since the amount of previous experience and the amount of contract money awarded are not associated with the quality of the evaluation, some other motivating force is required to upgrade laboratory improvement program evaluations.

3. It was shown that contractors adhere to the requirements of the funding agent's RFP's and tend not to exceed the basic minimum; therefore, it seems most logical and expedient for the

funding agent to be the motivating force to improve evaluations. This can be done by setting more explicit evaluation standards (such as those proposed in this research), incorporating them into the RFP's and enforcing them by being highly selective in the awarding of funds.

Further support for the last conclusion can be found in a General Accounting Office (GAO) evaluation of an educational experiment conducted under the auspices of the Office of Economic Opportunity (OEO) and reported in Cooley and Lohnes (1976, pp. 315-324). The experiment was a \$6 million educational innovation called performance contracting, which was believed to be capable of improving the reading and arithmetic skills of low income, low achieving public school children. "GAO as much as said that OEO wasted the \$6 million it expended on the study through misconception and mismanagement" (Cooley and Lohnes, 1976, p. 320). The GAO's criticisms were leveled against OEO not the individual contractors.

GAO contended that the true effects of the program could not be determined due to problems with the evaluation design and implementation. GAO noted that the experimental and control groups were not comparable. GAO blamed the OEO for not requiring the contractors to (1) monitor the performance of the control group, (2) collect adequate information on program effects, and (3) coordinate the length of the instructional periods so that valid comparisons could be made. The GAO also criticized OEO for not allowing the contractors sufficient lead-time to carefully plan and implement the program and its evaluation.

This report could very well foretell the outcome of an evaluation of the effectiveness of federally funded laboratory improvement programs, which have expended over \$4 million in the last three years. The circumstances surrounding the OEO-commissioned study parallel many of those discussed in this research.

Implications

For Laboratory Improvement Programs

For laboratory improvement programs, there are several relevant suggestions advanced by this research. The first is that in order for laboratory improvement programs to be effective, the real needs of the laboratory must be thoroughly defined. Performance deficiencies which are medically significant should be the focus of the program. The first step to improving performance is getting agreement that a problem exists (Fournies, 1978, p. 198). Krieg, Shearer, and Wenk (1978, p. 149) stress that performance deficiencies must be communicated in a specific, constructive, non-threatening manner. There may be a good explanation why certain laboratories perform tests beyond their skills. Clinicians may demand that the laboratory offer rare tests, for example. Laboratory improvement program staff must be sleuths first and then sages in order to solve problems. Unless there is an intense effort to uncover the real root of laboratory performance problems, and not just the superficial symptoms, legions of innovative new programs will be feeble change agents.

Program directors accountable to taxpayers must be bold enough to entertain the thought that some laboratory performance discrepancies are genuinely insignificant and do not warrant federal or state intervention. Precious financial resources must be conserved. Federally or state funded laboratory improvement programs must exhibit parsimony to win the confidence of the health care consumers and the health care professionals they hope to attract. Laboratory evaluators should be probing the real source of laboratory inaccuracy and the attendant threat to patient welfare. It is even conceivable that we (laboratory evaluators) have met the enemy, "and it is us." Evaluation measurement instruments often lack validity and reliability and yet evaluators persist in indicting laboratories for their unacceptable performance levels. For the moment, perhaps the most prudent expenditure of tax dollars would be an intensive research and development effort in devising appropriate clinical laboratory evaluation tools. Clearly, the knowledge and technology are available if not from the field of health, then from the fields of education and social science.

Only when the real performance problems are identified, determined to be important, and acknowledged as a problem by the laboratory, can the appropriate correction strategy be developed. Mager's and Pipe's (1970) algorithm for matching the problem to the solution should be followed.

Fournies (1978, pp. 195-201) pointed out that in business, 50 percent of unsatisfactory performance is related to feedback problems. The situation is analogous for laboratories. The

feedback that matters the most comes from within the organization: from clinicians, patients, and administrators. Inspection agents and proficiency test critiques are important but transient. To be effective, a laboratory improvement program must formulate goals consistent with the primary objectives of the clinical laboratory. Krieg, Shearer, and Wenk (1978, pp. 132-133) listed the following eleven laboratory goals:

1. Rapid turnaround time
2. Sufficient variety of services to meet requirements of the medical staff
3. Minimum number and severity of clinician complaints
4. Precision and accuracy as measured by quality control samples and proficiency test programs
5. Minimum coefficient of variation (error rate)
6. Minimum number of errors
7. High quality operation consistent with standards for good performance set by accreditation bodies
8. High staff morale as measured by turnover rates, questionnaires and informal conversations
9. Inservice training programs
10. Maximum productivity
11. Reasonable operating costs

An interdisciplinary task force with representation from the most influential groups can help formulate and promote worthwhile laboratory improvement program goals.

As soon as the goals and objectives have been set, the evaluation design should be selected and justified. If the program is to uncover its true impact on laboratory performance, the

director should consider a rigorous evaluation design. Criticisms against the use of experimental designs have been refuted well by Boruch (1976) who reviewed over 200 reports of social, medical and educational programs where rigorous experimental designs had been used. He concluded that the use of experimental design in real-world settings is feasible, cost effective, and ethical. Programs using experimental designs take advantage of, rather than neglect individual differences, and they foster worthwhile innovation. Boruch (p. 175) suggests that narrative and impressionistic information be used as an adjunct to rather than substitute for rigorous evaluation design. "With more experience in program evaluation, there is likely to be an increasing emphasis on the joint benefits of qualitative themes . . . coupled with systematic experimental research" (p. 175).

In evaluating laboratories or laboratory improvement programs, it would be advisable for evaluators to take on a more multivariate, multidimensional perspective. Since any one evaluation instrument has inherent flaws and limitations, it is better to use several, then find where the data converge and diverge or in Webb et al.'s phraseology, "find more points in conceptual space to triangulate" (1966). In the meantime, when proficiency test scores are reported to laboratories, to professional journals, to Congress and eventually to the general public, it would seem judicious to preface the report with a statement of the prevalence of the disease, and to be specific with regard to the kinds of diseases or conditions showing poor scores. Instead of a newspaper bannerhead that

reads "Rampant Laboratory Inaccuracy," it could read "Inability of Some Laboratories to Identify Rare Salmonella Species," or "Inability of Physicians Clinic Laboratories to Detect Unusual Red Blood Cell Antibody." This kind of specificity in reporting results of laboratory evaluations may facilitate more effective laboratory improvement efforts.

Perhaps if we better understood individual differences such as learning styles, motivation level, and work environment, we could better accommodate them in laboratory improvement programs. This would broaden the scope of current evaluation measures and may even enable program developers to improve some of these attributes (Messick, 1967).

For Evaluation as a Profession

There are implications of this research that extend beyond the parochial concerns of federally funded clinical laboratory improvement programs. Empirically based psychometric methods, such as factor analysis discussed in Chapter 2, could prove useful in consolidating and accurately classifying checklist items for future meta evaluation purposes. A larger data base than that used in this review would be required to afford optimum validity. Empirically based research exploring the interrelationships of individual evaluation factors, such as those incorporated into the checklist in Chapter 3, would be the next logical step in the evaluation of evaluations. The results of such endeavors would also provide useful information for further refining checklists to be used

prospectively by program directors and evaluators. An understanding of the intercorrelations among checklist items would clarify the construct of quality evaluation and may simplify the task of designing valid evaluations.

Stufflebeam (1978) suggests that research be extended to study the relationship between formative meta evaluation, which compares evaluation plans to procedural guidelines (as was done in Chapter 3), and summative meta evaluation, which sums up the overall merit of a completed evaluation (as was done in Chapter 4). The adherence to guidelines would be studied as the independent variables and the resulting quality of the finished evaluation (compared to established evaluation standards) would be the dependent variable (Stufflebeam, 1978). Such research would test the real payoff for adhering to evaluation guidelines, which at present is assumed to assure the quality of evaluation. The same approach could be used as in studies relating laboratory proficiency test performance to deficiencies uncovered during inspections (Peddecord, 1978). The process measures must correlate highly with the outcomes in order for guidelines and standards to be compelling regardless of whether it is the quality of laboratory performance or the quality of laboratory improvement programs under scrutiny.

The act of classifying evaluation activities as discussed in Chapter 3 can be seen as a step in the direction of a taxonomic perspective. Such a perspective can help to put order into the apparent chaos of program evaluation. The diversity and sheer bulk

of evaluation models and theories make good evaluation an onerous task for any but the most seasoned evaluator. Yet the need to be at least casually conversant with evaluation principles confronts almost everyone associated with a human service or educational program, due to growing demands for accountability and diminishing pecuniary resources.

The demand for trained evaluators may soon outstrip the supply (Worthen and Sanders, 1973, pp. 327-349). Who will be attracted to such an overwhelming vocation fraught with conflicting theories and lacking a clear sense of direction? The time is ripe for a new breed of evaluation systematists who can organize the body of evaluation knowledge as adroitly as Bloom (1956), Krathwohl (1964), and Harrow (1972) sorted behavioral terms by the domains and levels of learning. Just as their taxonomies "provided a common foundation upon which teachers could organize learning experiences . . . [and] enabled professionals to communicate" (Harrow, 1972, p. 9), taxonomies for the domains of program evaluation would enable evaluators to plan better evaluations. A common evaluation vernacular would unite health, human service and educational professionals in their efforts to provide quality, cost effective programs.

There are a few individuals who have sensed this need. William Gephart (n.d.) developed a very readable structure for synthesizing models of evaluation. Rossi and his group (1979) published a text to engender a systematic approach to evaluation. It is interesting that the book was inspired by a meeting where all

three authors presented independently written but remarkably similar papers (p. 11). Apparently, there is hope. Evaluation ideas are coming together; consensus is possible. What is left to do now is to transmit the evaluation knowledge base to the lay individuals who most need the information. Equipped with a working understanding, program staff can take care of the evaluation preliminaries, select qualified consultants, and put valid, reliable evaluations to work for the benefit of their programs and their clients.

APPENDIX

Below is an alphabetical listing of the twenty-three laboratory improvement program proposals. They do not appear in the same order as listed in Table 3. The anonymity of individual contractors has been preserved with regard to the technical adequacy ratings discussed in this thesis.

Colorado Association for Continuing Medical Laboratory Education Training Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Commonwealth of Kentucky Laboratory Improvement Program Proficiency Testing Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Commonwealth of Kentucky Laboratory Improvement Program Technical Consultation Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Connecticut State Department of Health Proficiency Testing Proposal, 1979.

Idaho State Health Department Training Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Iowa State Hygienic Laboratory Proficiency Testing Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Kansas Department of Health and Environment Technical Consultation Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Massachusetts Health Research Institute, Inc. Training Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Massachusetts Health Research Institute, Inc. Proficiency Testing Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Michigan Department of Public Health Training Proposal, 1979.

Minnesota Department of Health Proficiency Testing Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Missouri Division of Health Technical Consultation Proposal and Technical Portion of the Best and Final Business Offer, 1979.

North Carolina Department of Human Resources Technical Consultation Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Ohio State University Training Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Rhode Island Department of Health Technical Consultation Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Rhode Island Training Proposal and Technical Portion of the Best and Final Business Offer, 1979.

South Carolina Division of Laboratory Improvement Training Proposal and Technical Portion of the Best and Final Business Offer, 1979.

South Dakota Department of Health Technical Consultation Proposal, 1979.

South Dakota State Health Department (Subcontracted to the University of South Dakota) Training Proposal, 1979.

University of North Dakota Training Proposal and Technical Portion of the Best and Final Business Offer, 1979.

University of Wisconsin State Laboratory of Hygiene Proficiency Testing Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Utah State Health Department Training Proposal and Technical Portion of the Best and Final Business Offer, 1979.

Wisconsin Division of Health Technical Consultation Proposal and Technical Portion of the Best and Final Business Offer, 1979.

REFERENCES

- Alkin, M. C., R. Daillak, and P. White (1979) Using evaluations: Does evaluation make a difference. Beverly Hills: Sage.
- "An analysis of a laboratory improvement program for Idaho private physician office laboratories through technical consultation" (1979a) Final report of contract #200-77-0742 to the Center for Disease Control, Bureau of Laboratories, Atlanta, Georgia. (unpublished)
- "An analysis of Idaho private physician office laboratory facilities and testing activities through on-site inspection" (1979b) Final report of contract #200-77-0716 to the Center for Disease Control, Bureau of Laboratories, Atlanta, Georgia. (unpublished)
- Anderson, S. B., and S. Ball (1978) The profession and practice of program evaluation. San Francisco: Jossey-Bass.
- Bartz, A. E. (1976) Basic statistical concepts in education and the behavioral sciences. Minneapolis: Burgess.
- Belk, W. P., and F. W. Sunderman (1947) "A survey of the accuracy of chemical analysis in clinical laboratories." American Journal of Clinical Pathology 17:853-861.
- Benson, J., and L. Crocker (1979) "The effects of item format and reading ability on objective test performance: A question of validity." Educational and Psychological Measurement 39:381-387.
- Berg, A. O. (1979) "Does continuing medical education improve the quality of medical care? A look at the evidence." Journal of Family Practice 8:1171-1174.
- Black, W. A., S. E. Dorse, and J. L. Whitby (1976) "A regional quality control program in microbiology. Parts I and II." American Journal of Clinical Pathology 66:401-415.
- Bloom, B. S. [ed.] (1956) Taxonomy of educational objectives handbook I: Cognitive domain. New York: David McKay.
- Borg, W. R., and M. D. Gall (1979) Educational research: An introduction, 3rd ed. New York: Longman.

- Boruch, R. F. (1976) "On common contentions about randomized field experiments," in G. V. Glass (ed.) Evaluation studies review annual, Volume 1. Beverly Hills: Sage.
- Brooks, M. L. (n.d.a) Primer for workshop leaders: A guide for laboratory trainers. Atlanta: U.S. Department of Health, Education and Welfare, Center for Disease Control.
- _____ (n.d.b) Analyzing needs: A primer for trainers. Atlanta: U.S. Department of Health Education and Welfare, Public Health Service, Center for Disease Control.
- Brown, C. R. (1977) "The continuing education component of the bicycle approach to quality assurance," in R. H. Egdahl and P. M. Gertman (eds.) Quality health care: The role of continuing medical education. Germantown, Maryland: Aspen Systems Corporation.
- Campbell, D. T. (1967) "Administrative experimentation, institutional records, and nonreactive measures," in J. C. Stanley (ed.) Improving experimental design and statistical analysis. Chicago: Rand McNally.
- _____ (1969) "Factors relevant to the validity of experiments in social settings," in H. C. Schulberg, A. Sheldon, and F. Baker (eds.) Program evaluation in the health fields. New York: Human Sciences Press.
- _____ and J. D. Stanley (1963) Experimental and quasi-experimental designs for research. Chicago: Rand McNally.
- Caplan, R. M. (1973) "Measuring the effectiveness of continuing medical education." Journal of Medical Education 48: 1150-1152.
- Carlson, D. J. (1977) "Cost effectiveness of laboratory improvement programs: The viewpoint from the private sector." Health Laboratory Science 14, 3: 199-205.
- Carroll, M. R. (1980) "Social structure among proposers of 'models' of the evaluation process in education." Paper presented at the annual meeting of the American Educational Research Association, Boston, Massachusetts, April, 1980.
- Casscells, W., A. Schoenberger, and T. B. Graboys (1978) "Interpretation by physicians of clinical laboratory results." New England Journal of Medicine 299, 18: 999-1000.

- Center for Disease Control: Requests for proposals (1979)
200-79-0911(P), 200-79-0912(P), and 200-79-0913(P), technical
proposal instructions, section B-2.
- Clinical Laboratory Improvement Act of 1967, P.L. 90-174. Code
of federal regulations Title 42, part 74. Washington: U.S.
Department of Health Education and Welfare.
- College of American Pathologists (1974) Standards for accreditation
of medical laboratories. Chicago: Commission on Inspection
and Accreditation of the College of American Pathologists.
- Connelly, T. (1979) "Continuing education in allied health: The
state of the art." Journal of Allied health 8, 1: 38-45.
- Cook, T. D., and D. T. Campbell (1976) "The design and conduct of
quasi-experiments and true experiments in field settings,"
pp. 223-326, in M. D. Dunnette (ed.) Handbook of industrial and
organizational psychology. Chicago: Rand McNally.
- Cooley, W. W., and P. R. Lohnes (1976) Evaluation research in
education. New York: Irvington.
- Davidge, A. M., W. K. Davis, and A. L. Hull (1980) "A System for
the evaluation of medical students clinical competence."
Journal of Medical Education 55: 65-67.
- Dielman, T. E., T. E. Hull, and W. K. Davis (1980) "Psychometric
properties of clinical performance ratings." Evaluation and
the Health Professions 3: 103-117.
- Dixon, J. (1978) "Evaluation criteria in studies of continuing
education in the health professions: A critical review and
suggested strategy." Evaluation and the Health Professions
1: 47-65.
- "Do medical laboratories need tighter control" (1979) Salt Lake
City, Deseret News, January 8, 1979, A5, col. 1.
- Donabedian, A. (1969) "Evaluating the quality of medical care,"
pp. 186-215, in H. C. Schulberg, A. Sheldon, and F. Baker (eds.)
Program evaluation in the health fields. New York: Human
Sciences Press.
- "The effect of PSRO's on health care costs: Current findings and
future evaluations" (1979) Washington, D.C.: U.S. Government
Printing Office, The Congress of the U.S. Congressional Budget
Office.
- Engel, J. D. (1978) "Validation of domain referenced test items."
Evaluation and the Health Professions 1: 111-119.

- Evaluation Research Society (1980) Standards for program evaluation. Exposure draft, May, 1980, Brooklyn, New York.
- Fifer, W. R. (1979) "Quality assurance: Debate persists on goals, impact and methods of evaluating care." Hospital, Journal of the American Hospital Association April 1: 163-167.
- "Final report of the technical consultation program under Contract no. 200-77-0743." (1979) Utah State Division of Health, Bureau of Laboratories, to the Center for Disease Control, Bureau of Laboratories, Atlanta, Georgia.
- Fink, A., and J. Kosecoff (1978) An evaluation primer. Washington, D.C.: Capitol Publications.
- Finkel, P. W., and T. R. Miller (1973) "A proficiency test assessment of clinical laboratory capability in the U.S." A report prepared for the Division of Health Evaluation, Department of Health, Education and Welfare, NBSIR 73-163. Washington, D.C.
- Forney, J. E., and M. M. Brooke (1967) "Role of the public health service in the improvement of clinical laboratories." Health Laboratory Science 4, 2: 62-69.
- _____, J. M. Blumberg, M. M. Brooke, E. Eavenson, R. K. Gilbert, and W. Kaufmann (1978) "Laboratory evaluation and certification," pp. 127-171, in S. L. Inhorn (ed.) Quality assurance practices for health laboratories. Washington, D.C.: American Public Health Association.
- Fournies, F. F. (1978) Coaching for improved work performance. New York: Van Nostrand and Reinhold.
- Fouty, R. A., V. E. Haggren, and J. D. Sattler (1974) "Problems, personnel, and proficiency of small hospital laboratories." Public Health Reports 89, 5: 408-417.
- Freeman, H. E., and C. C. Sherwood (1969) "Research in large-scale intervention programs," pp. 73-91, in H. C. Schulberg, A. Sheldon, and F. Baker (eds.) Program evaluation in the health fields. New York: Human Sciences Press.
- Fromkin, H. L., and S. Streuffert (1976) "Laboratory experimentation," pp. 415-465, in M. D. Dunnette (eds.) Handbook of industrial and organizational psychology. Chicago: Rand McNally.
- Garcia, K. W. (1980) Preparing for a laboratory inspection [workshop manual]. Montana State Health Department, Laboratory Improvement Program, Helena, Montana.

- Gephart, W. J. (n.d.) Evaluation: Past, present and future. Occasional paper 17. Bloomington, Indiana: Phi Delta Kappa.
- Glass, G. E. (1977) "Integrating findings: The meta analysis of research," in L. S. Shulman (ed.) Review of research in education. Volume 5. Itasca, Ill.: F. E. Peacock.
- Gronlund, N. E. (1976) Measurement and evaluation in teaching. 3rd ed. New York: Macmillan.
- Hamilton, D. (1977) "Making sense of curriculum evaluations: Continuities and discontinuities in an evaluation idea," in L. S. Shulman (ed.), Review of research in education. Volume 5. Itasca, Ill.: F. E. Peacock.
- Hamilton, J. A., O. V. Baker, and A. M. Mitchell. (1979) "Identifying well-evaluated activities in career education." Measurement and evaluation in guidance 12: 116-120.
- Harasym, P. H., D. A. Norris, and F. L. Lorscheider (1980) "Evaluating student multiple-choice responses: Effects of coded and free formats." Evaluation and the Health Professions 3, 1: 63-84.
- Hardison, J. E. (1979) "To be complete." New England Journal of Medicine 300, 4: 193-194.
- Harrow, A. J. (1972) A taxonomy of the psychomotor domain. New York: David McKay.
- Hayman, J. L., and R. N. Napier (1975) Evaluation in the schools: A human process for renewal. California: Wadsworth.
- Homans, G. C. (1978) Sentiments and activities: Essays in social science. New York: Free Press of Glencoe, 1962, Cited by L. M. Smith, "An evolving logic of participant observation, educational ethnography and other case studies," in L. S. Schulman (ed.) Review of research in education. Volume 6. Itasca, Ill.: F. E. Peacock.
- Inui, T. S., E. L. Yourtee, and J. W. Williamson (1976) "Improved outcomes in hypertension after physician tutorials: A controlled trial." Annals of Internal Medicine 84: 646-651.
- Jaeger, R. M. (1978) "About educational indicators: Statistics on the conditions and trends in education," in L. S. Shulman (ed.) Review of research in education. Volume 6. Itasca, Ill.: F. E. Peacock.

- Javits, J. K. (1979) [Personal correspondence between Senator Jacob K. Javits and HEW Secretary Patricia Roberts Harris]. Reprinted with permission in D. W. Weissman (ed.) National intelligence report: Clinical labs/blood banks, December 18, 1979, 1, 5: 2.
- Jessee, W. F., W. B. Munier, J. E. Fielding, and M. J. Goran (1975) "PSRO: An educational force for improving quality of care." The New England Journal of Medicine 292: 668-674.
- Joint Commission on the Accreditation of Hospitals (1976) Accreditation manual for hospitals. Chicago: Joint Commission on the Accreditation of Hospitals.
- Kassirer, J. P., and S. G. Pauker (1978) "Should diagnostic testing be regulated?" New England Journal of Medicine 299: 947-949.
- Kauffman, N. M. (1979) "Clinical laboratory improvement legislation: An analysis." American Journal of Medical Technology 45: 9: 813-815.
- Kaufmann, W. (1973) "Quality control of physicians office laboratories." Health Laboratory Science 10, 4: 284-286.
- Krathwohl, D. R., B. S. Bloom, and B. B. Masia (1964) Taxonomy of educational objectives handbook II: Affective domain. New York: David McKay.
- Krieg, A. F., L. K. Shearer, and R. E. Wenk (1978) Laboratory communication: Getting your message through. Oradell, New Jersey: Medical Economics.
- Kukuk, C. R., and C. F. Baty (1979) "The misuse of multiple regression with composite scales obtained from factor scores." Educational and Psychological Measurement 39: 277-290.
- Kull, D. J. (1980) "State Licensure laws for Laboratorians." Medical Laboratory Observer 12, 1: 72-105.
- "Laboratory proficiency" (1976) An editorial. British Medical Journal 1, 6000: 5.
- La Motte, L. C. (1977) "The impact of laboratory improvement programs on laboratory performance: The CLIA 67 experience." Health Laboratory Science 14:213-223.
- _____, G. O. Guerrant, D. S. Lewis, and C. T. Hall (1977) "Comparison of laboratory performance with blind and mail-distributed proficiency testing samples." Public Health Reports 92: 554-560.

- Lloyd, J. S., and S. Abrahamson (1979) "Effectiveness of continuing medical education: A review of the evidence." *Evaluation and the Health Professions*, 2: 251-280.
- Loughmiller, G. C., R. L. Ellison, C. W., Taylor, and P. B. Price (1970) "Predicting career performance of physicians using the biographical inventory approach." *Proceedings, 78th Annual Convention of the American Psychological Association*, 5: 153-154.
- Luft, A. S., J. P. Bunker, and A. C. Enthoven (1979) "Should operations be regionalized?" *New England Journal of Medicine* 201, 25: 1364-1369.
- Lyons-Morris, L., and C. Taylor-Fitz-Gibbon (1978) *Evaluator's handbook*. Beverly Hills: Sage.
- Mager, R. K., and P. Pipe (1970) *Analyzing performance problems*. Belmont, Calif.: Fearon Pitman.
- "Manpower for the medical laboratory" (1967) A report of a conference of government and the professions. Washington, D.C.: U.S. DHEW Public Health Service, PHS #1833 and 1771.
- Mantel, N., and W. Haenszel (1959) "Statistical aspects of the analysis of data from retrospective studies." *Journal of the National Cancer Institute* 22: 719-748.
- Messick, S. (1967) The criterion problem in the evaluation of instruction: Assessing possible not just intended outcomes. *Proceedings of the Symposium on Problems in the Evaluation of Instruction*. UCLA: CSE Report No. 22.
- McCormick, W., J. A. Ingelfinger, G. Isakson, and P. Goldman (1978) "Errors in measuring drug concentrations." *The New England Journal of Medicine* 299: 1118-1121.
- McDonald, C. J. (1974) "Protocol based computer reminders, the quality of care and the nonperfectability of man." *New England Journal of Medicine* 295: 1351-1355.
- McGuckin, M. B., A. F. Adenbaum, and E. Corbin (1979) "Abnormal results are ignored by physicians." *Lab World* 30, 12: 29-30.
- McGuire, C., R. E. Hurley, D. E. Babbott, and J. S. Butterworth (1964) "Auscultatory skill: Gain and retention after intensive instruction." *Journal of Medical Education* 39: 120-131.
- Newble, D. I., A. Baxter, and R. G. Elmslie (1979) "A comparison of multiple choice tests and free-response tests in examinations of clinical competence." *Medical Education* 13: 263-268.

- Newstrom, J. W. (1978) "Catch--22: The problems of incomplete evaluation of training." *Training and Development Journal* 32, 11: 22-24.
- Notice of proposed rulemaking (1979) Personnel standards for clinical laboratories, Document #79, 31647. *Federal Register*, Washington, D.C., 58923-58928.
- Nunnally, J. C. (1978) *Psychometric theory*. 2nd ed. New York: McGraw Hill.
- Olson, R. P., and M. F. Fruin (1979) "Evaluation doesn't have to be difficult." *Journal of Extension* 17: 21-25.
- Page, G. G., A. D. Van Wart, D. E. Raudzus, and D. M. Kettys (1979) "The effect of continuing medical education programs on clinical practice: Fact or fantasy." *Medical Education* 13: 292-297.
- Peddecord, K. M. (1978) *Clinical laboratory proficiency test performance: Its relationship to environmental, structural and process variables*. Doctoral dissertation, University of Texas, School of Public Health, Houston.
- Pierleoni, R. G. (1978) "Clinical evaluation techniques for the health professions." *Improving Human Performance Quarterly* 7: 204-216.
- Posavac, E. J. (1980) "Evaluation of patient education programs: A meta analysis." *Evaluation and the Health Professions* 3: 47-62.
- Reed, D. E., C. Lapeñas, and K. D. Rogers (1973) "Continuing education based on record audit in a community hospital." *Journal of Medical Education* 48: 1152-1155.
- Relman, A. S. (1979) "Technology costs and evaluation." *New England Journal of Medicine* 301: 1444-1445.
- Rosenthal, R., and R. L. Rosnow (1975) *The volunteer subject*. New York: John Wiley and Sons.
- Rossi, P. H., H. E. Freeman, and S. R. Wright (1979) *Evaluation: A systematic approach*. Beverly Hills: Sage.
- Sattler, J. (1970) "Continuing education of laboratory personnel." *American Journal of Medical Technology* 36, 5: 239-243.
- Schaeffer, M., D. Widelock, S. Blatt, and M. E. Wilson (1967) "The clinical laboratory improvement program in New York City: I. Methods of evaluation and results of performance tests." *Health Laboratory Science* 4, 2: 72-89.

- _____, P. S. May, S. Blatt, and M. E. Wilson (1970) "The clinical laboratory improvement program in New York City: II. Progress after five years of experience." *Health Laboratory Science* 7, 4: 242-255.
- Schoen, I., G. D. Thomas, and S. Lange (1971) "The quality of performance in physicians office laboratories." *American Journal of Clinical Pathology* 55: 163-169.
- Scriven, M. (1974) "Evaluation perspectives and procedures," pp. 3-93, in W. J. Popham (ed.) *Evaluation in education*. Berkeley: McCutchan.
- _____. (1976) "Evaluation bias and its control," in G. V. Glass (ed.) *Evaluation studies review annual*. Volume I. Beverly Hills: Sage.
- Sealfon, M. S. (1976) "Definitions, sources, and detection of laboratory error, a review." *American Journal of Medical Technology* 42: 476-480.
- Shepard, L. A. (1976) A checklist for evaluating large-scale assessment programs. Boulder, Colorado: University of Colorado. ERIC Document ED 163 057.
- Sherman, C. (1979) "And it turned out the lab made a mistake." *Prevention*, April: 81-85.
- Siegel, S. (1956) *Nonparametric statistics for the behavioral sciences*. New York: McGraw Hill.
- Simpson, W. J. (1979) "Practice monitoring as a means to direct individual continuing medical education." *Southern Medical Journal* 72: 852-853.
- Smith, L. M. (1978) "An evolving logic of participant observation, educational ethnography and other case studies," in L. S. Shulman (ed.) *Review of research in education*. Volume 6. Itasca, Ill.: F. E. Peacock.
- Smith, M. L., and G. V. Glass (1977) "Meta analysis of psychotherapy outcome studies." *American Psychologist* 32: 752-760.
- Steele, B. W., M. K. Schauble, J. M. Becketl, and J. E. Bearman (1977) "Evaluation of clinical chemistry laboratory performance in twenty veterans administration hospitals." *American Journal of Clinical Pathology* 67: 594-602.
- Stufflebeam, D. L. (1978) "Meta evaluation: An overview." *Evaluation and the Health Professions* 1: 17-43.

- Taylor, C. W., P. B. Price, J. M. Richards, and T. L. Jacobsen (1964) "An investigation of the criterion problem for medical school faculty." *Journal of Applied Psychology* 48: 294-301.
- _____, and J. J. Richards (1965) "An investigation of the criterion problem for a group of surgical general practitioners." *Journal of Applied Psychology* 49: 399-406.
- Taylor Fitz-Gibbon, C., and L. Lyons-Morris (1978a) *How to design a program evaluation*. Beverly Hills: Sage.
- _____. (1978b) *How to calculate statistics*. Beverly Hills: Sage.
- "U nursing study to determine impact of continuing education" (1979) University of Utah Health Sciences Report, January: 3.
- U.S. Congress (1977) Senate committee on human resources, subcommittee on health and scientific research. *Clinical Laboratory Improvement Act of 1977*. Hearings, 95th Cong., on S705, March 29 and 30, Washington, D.C.: Government Printing Office.
- U.S. Department of Health, Education, and Welfare (1978) *Health Care Financing Administration's Health Standards and Quality Bureau. Office of Standards and Certification. Clinical Laboratory Guidelines Medicare*. Washington, D.C.: Government Printing Office.
- Wallace, M. [1970] CBS "60 Minutes" [T.V.] report on fraud in medical laboratories in Chicago. Don Hewitt, Executive Producer. New York City.
- Webb, E. J., D. T. Campbell, R. D. Schwartz, and L. Sechrest (1966) *Unobtrusive measures. Nonreactive research in the social sciences*. Chicago: Rand-McNally.
- White, W. D. (1979) *Public health and private gain*. Chicago: Maaroufa Press.
- Wigton, R. S. (1980) "Factors important in the evaluation of clinical performance of internal medicine residents." *Journal of Medical Education* 55: 206-208.
- Williamson, J. W., M. Alexander, and G. E. Miller (1967) "Continuing education and patient care research: Physician response to screening test results." *Journal of the American Medical Association* 201: 118-122.
- Worthen, B. R., and J. R. Sanders (1973) *Educational evaluation: Theory and practice*. Belmont, Calif.: Wadsworth.

VITA

Name	Deborah Joan Clarke del Junco (Illes)
Birthdate	June 10, 1951
Birthplace	Chicago, Illinois
Education	Western Illinois University Macomb, Illinois B.S. in Medical Technology 1970-1974 Bowling Green State University Bowling Green, Ohio Honors Program 1969-1970 Highland Park High School Illinois 1967-1969
Professional Certification	Medical Technologist, American Society of Clinical Pathologists, #095550 Clinical Laboratory Scientist, National Certification Agency for Medical Laboratory Personnel, #783997-5
Professional Positions	Training Coordinator, Utah State Health Laboratory, Salt Lake City, Utah, 1977 to 1980; Microbiologist, LDS Hospital Infectious Disease Laboratory, Salt Lake City, 1977 to 1979; Microbiologist, Univer- sity of Utah Medical Center, Clinical Laboratory, Salt Lake City, 1976-1977; Microbiologist/ Infection Control Coordinator, Payson Hospital, Payson, Utah, 1975-1976; Hematology Technologist, LDS Hospital Laboratory, Salt Lake City, Utah, 1974-1975

Professional
Organizations

American Society for Microbiology, American Public Health Association, American Society for Medical Technology, American Society of Clinical Pathologists, American Society of Allied Health Professions, American Association for the Advancement of Science, Evaluation Research Society, Colorado Association for Continuing Medical Laboratory Education.

Publications

del Junco, D. J., B. Gardner, and J. H. Hengesbaugh (1980) An educational approach to laboratory improvement. Paper accepted for presentation, American Public Health Association (annual meeting), October 20, 1980

del Junco, D. J., J. Clayton, B. K. Hudson and M. R. Britt (1979) Determining quality of routine bacteriology--An alternative approach. Paper accepted for presentation, American Society for Microbiology (annual meeting), May 7, 1979.