EVALUATION OF TRANSFER OF TECHNICAL TRAINING:
A PROTOTYPE

THESIS

Presented to the Graduate Council of the University of North Texas in Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

By

Carol A. Stubbs, B.S.
Denton, Texas
August, 1993

The degree of transfer of technical training to workplace behavior was evaluated using a Solomon Four-Group experimental design. Additionally, all groups received retrospective pretests. Subjects were 103 technicians in an electronics company. Supervisors rated technicians on behaviorally anchored rating scales which were developed and labeled as behavior description scales for simplicity. Analysis of variance revealed no effect for training nor pretest. A training-pretest interaction effect was revealed for one dimension (Communication with Support Groups). Analysis of covariance revealed main effects for pretesting for two dimensions (Problem Solving and Communication with Supervisor) and a pretest-training interaction for one dimension (Problem Solving). Except for one dimension, t tests revealed no significant differences between traditional pretests and retrospective pretests, thus negating a hypothesized response shift bias.
ACKNOWLEDGEMENTS

The researcher would like to thank the thesis committee, Dr. Douglas Johnson, Dr. Michael Beyerlein, and Dr. Jerry Bayless, for their effort, expertise, encouragement, and patience. In addition, the researcher would like to thank Dr. Richard Gunst for his time and enthusiasm in imparting his knowledge and expertise to make statistics comprehensible to the business sponsors of the project and Dr. Mary Alice Gordon, who made herself available to encourage and offer advice in transcending the complexities of the business world in order to carry out a scientific study.
TABLE OF CONTENTS

LIST OF TABLES ........................................ iv
LIST OF ILLUSTRATIONS ................................. v

Chapter

I. INTRODUCTION TO THE STUDY ..................... 1

Approaches to evaluating training
Quantitative versus qualitative approach
Formative versus summative approaches
Results-oriented approaches
Quality control Approach
The Importance of Needs Assessment
Levels of Evaluation
Kirkpatrick’s four levels of evaluation
Level one evaluation: Reaction
Level two evaluation: Learning
Level three evaluation: Behavior
Level four evaluation: Results
Critique of Kirkpatrick’s model
Issues in Performance Rating
Behaviorally Anchored Rating Scales
Retrospective Pretests as a Method
Experimental Design and Statistical Analysis

II. METHOD .............................................. 66

Subjects
Demographic variables
Materials
Dependent variables
Independent variables
Procedure
Development of behaviorally anchored rating scales
Administration of evaluation instrument

III. RESULTS ........................................... 77
Hypothesis 1
Hypothesis 2
Hypothesis 3
TABLE OF CONTENTS--Continued

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV. DISCUSSION</td>
<td>84</td>
</tr>
<tr>
<td>Future Research</td>
<td></td>
</tr>
<tr>
<td>APPENDICES</td>
<td>93</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>118</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Posttest Means and Standard Deviations for Dimensions 1 - 8</td>
<td>106</td>
</tr>
<tr>
<td>2. Paired t test on Pretest and Retrospective Pretest</td>
<td>109</td>
</tr>
<tr>
<td>3. Analysis of Variance Results with Significance</td>
<td>110</td>
</tr>
<tr>
<td>4. Analysis of Covariance Results with Significance (using Retrospective Pretest and Worksites as Covariates)</td>
<td>112</td>
</tr>
<tr>
<td>5. Testing Group Effect of Pretested Individuals</td>
<td>114</td>
</tr>
<tr>
<td>6. Correlation Matrices for Pretest, Posttest, and Retrospective Pretest Scores</td>
<td>115</td>
</tr>
</tbody>
</table>
LIST OF ILLUSTRATIONS

Figure                                      Page

1. Theoretical Model for Determining Content Validity       6

2. Pretest-Treatment Interaction for Dimension 5
   (Communication with Support Group) Based on ANOVA       79

3. Pretest-Treatment Interaction Effect for Dimension 1
   (Problem Solving) based on Analysis of Covariance       80
CHAPTER I

INTRODUCTION TO THE STUDY

Traditionally, training evaluation has been centered upon the quality of training methods. Proof of effectiveness was demonstrated when trainees were able to score well on knowledge tests administered at the end of the training. The economic market of the 1990's has wrought changes in perceived needs. The attempts of organizations to survive globally competitive markets has implanted the demand for all business functions in corporations to be justified. The training function must answer questions applied across all departments. Not only must training departments assure managers of effective training methods, but also that training adds value to the business process. Evaluation of training, which in the past was designed to please trainers and trainees, must now answer questions related to improving total organizational effectiveness.

The present study was instituted to address the issue of transfer of technical training to workplace behaviors. Trainers in an electronics manufacturing corporation were interested in knowing if concepts taught in Physics/Chemistry classes could be proven to increase performance for process and engineering technicians.
First in a series of classes designed to impart knowledge of technical operations in the components sector of an electronics company, Physics/Chemistry classes are designed in three phases. Many technicians take only the one Physics/Chemistry class most appropriate to understanding their operation. Others take two of the long-existing classes. Others take a third class, only recently developed. The classes cover concepts chiefly, with a minimum of calculations. The concepts undergird operations in various operational units.

Physics/Chemistry classes are the most remote of all technical training from actual work behaviors. They form a basis for understanding later training in the series such as Chemical Vapor Deposition, Plasma Etching, and other courses which teach about specific work processes and how to implement them. Physics/Chemistry classes were chosen for evaluation by training managers in the Components Sector Training and Organizational Effectiveness organization in the hope that success with a class remote from work behaviors would assure success in more behavior-oriented classes.

Approaches to Evaluating Training

An essential element of understanding any approach to training evaluation depends upon a knowledge of the evaluation taxonomy developed by Kirkpatrick (Birnbrauer, 1990) and used almost universally by training experts.
Level One evaluation, as explained by Kirkpatrick's model, evaluates the trainees' reactions. A measure is taken of trainees' satisfaction with the training. Level Two evaluation measures actual learning. Usually, a traditional paper-and-pencil test provides the measure. Level Three of Kirkpatrick's model evaluates transfer of training and seeks to analyze the extent to which training is demonstrated in job behaviors. Level Four evaluation seeks answers to how much training contributes to results for the corporation. Kirkpatrick's four levels are often summarized as: (1) reaction, (2) learning, (3) behavior, and (4) results.

Quantitative versus Qualitative Approaches

While urging scientific inquiry tempered by business contingencies, Goldstein (1986) pointed out several questions that need to be answered by evaluation of training. A first concern was whether various criteria indicated that change had occurred; further, whether the changes could be attributed to the instructional program. Evaluators would also need to know whether changes could be replicated for new participants in the same program. A final question would be whether new participants would experience similar changes in the same program in a new organization. These questions would need to be asked at each criterion level (reaction, learning, behavior, and results).
The evaluation process can be viewed in terms of three evolutionary phases, according to Goldstein (1986). He referred to the two most primitive representations of the phases as negativists and positivists. He described negativists as those who think that training in organizations can not be evaluated scientifically. Negativists would base decisions upon anecdotal trainee-trainer relations. A later evaluation phase was represented by positivists who think that training must be evaluated with scientific rigor as though in a laboratory. Positivists, as he described them, were prone to ignore organizational and environmental constraints. Under such thinking, studies might fail to be done where the sample was too small for traditional designs.

Goldstein's (1986) description of the latest representation of evolutionary phases was the activist. The activist seeks the most systematic procedures possible to fit the existing situation. He or she controls as many of the extraneous variables as possible, while recognizing the limitations of the design. Goldstein claimed that even with an uncontrollable environment, knowledge of important factors in experimental design would make evaluation useful.

Content validity could be a starting point of evaluation, according to Goldstein (1986). Evaluation based on an appropriate needs assessment and reflecting
knowledge, skills, and abilities would be judged as having content validity. Goldstein presented a table with four quadrants representing key concerns. The horizontal axis of the table represented the importance of KSAs, while the vertical axis was used to show degree of emphasis in training.

In Goldstein's (1986) fourfold table, Quadrant A depicted a situation wherein unimportant knowledge, skills, and abilities (KSAs) were not emphasized in training. Quadrant B showed the situation of important KSAs not being emphasized. Quadrant C of the model regarded the idea that non-important KSAs might be emphasized. Quadrant D represented important KSAs being emphasized in training. By having subject matter experts (SMEs) in the job categorize job tasks and SMEs in training categorize concepts taught in training, an examiner could determine clearly the content validity of a training program.

One model of evaluation which Goldstein (1986) described was the "individual differences model." The individual differences model correlates training test scores with job performance evaluations or other job-related criterion measures. Addressing the individual differences approach, several researchers (Kraut, & Gordon & Cohen, cited in Goldstein, 1986) found level of success
Figure 1. Theoretical Model for Determining Content Validity

<table>
<thead>
<tr>
<th>Importance of KSA (as determined by Needs Assessment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Important</td>
</tr>
<tr>
<td>Not Emphasized</td>
</tr>
<tr>
<td>Degree of Emphasis in Training</td>
</tr>
<tr>
<td>Emphasized</td>
</tr>
<tr>
<td>+ C</td>
</tr>
</tbody>
</table>

on early training tasks to predict level of success on later training tasks and level of success on training to predict future promotions and job performance evaluations.

The problem with the individual difference model seems to be that people who do well on training test scores tend to do well on job performance factors as well. No proof is given that the class covered knowledge, skills, or abilities (KSAs) needed for the job or that learning took place. To remediate the lack of definitive information, Goldstein advocated addition of experimental designs and control groups. A thorough needs assessment, according to Goldstein, also lends credence to the approach by assuring that training addresses the required KSAs. According to Goldstein, the individual differences
model is actually better for selection than for training evaluation.

Brinkerhoff (1989) made a rather weak case for abnegating all efforts to prove training effectiveness through experimental method. He stated, "...training is often a necessary given because it is vital to the successful implementation of organizational goals. In such instances, it is the role of evaluation to ensure that training makes the most effective and efficient contribution possible, rather than to question how much contribution it alone has made." He argued that a solid needs assessment was needed first, and not an experimental-design-based evaluation. Then evaluation should measure effective and efficient delivery of training and contribution of training to organizational goals.

Certain training efforts appear to call for results that are not immediately obvious. McEvoy and Buller (1990) discussed the situation wherein training is a perquisite and therefore not necessarily appropriate to evaluate beyond the level of satisfaction. They mentioned outdoor management training as belonging to that genre of training, possibly carrying with it rewarding experiences which build team spirit or other undefined, but desirable, outcomes. Orientation, power enhancement, and small-scale
organizational development were also mentioned as non-measurable purposes of training.

Training exists for many reasons which are not immediately obvious or measurable, argued McEvoy and Buller (1990). They lamented the possibility of hard-measure proponents who demand quantitative, measurable improvements in behavior driving certain qualitatively valuable training out of the market.

Formative versus Summative Approaches

To set a purpose for evaluation one must decide whether the evaluation is formative or summative, according to Ratzlaff (cited in Birnbrauer, 1990). Formative is another word for improvement-oriented. A formative design is used to decide whether a program should be modified, raising questions about how to staff the program and select participants. A formative evaluation assesses the success of each individual stage and component. In contrast, summative evaluation determines whether a program should be expanded, maintained, contracted out, or eliminated. Summative evaluation seeks to characterize the participants who benefited most from the training and to assesses the completed program as a unit. The formative approach seems to be most common.

Selection of method and design is driven by the fact that someone is paying for the training. Hence, the
trainer must find out how the client who pays for the training wants to use the evaluation information. The trainer can then use information about who is purchasing training to decide what data will be collected and how findings will be presented.

A formative strategy is recommended by Bushnell (1990). Bushnell’s four-step process consists of: (1) identifying evaluation goals, (2) developing an evaluation design and strategy, (3) analyzing the data, and (4) reaching conclusions and making recommendations.

Bushnell (1990) divides the second phase, developing an evaluation design and strategy, into four parts, that of: (1) selecting and/or constructing appropriate measuring tools, (2) matching data types with experimental designs, (3) allocating resources to data collection, and (4) identifying appropriate data sources. The third, or analysis, step ties the gathered data to the original goals of the evaluation. The trainer must determine whether the data are pertinent and whether they are sufficient. One must also consider the disruptiveness of the collection process and appropriateness of analytical procedures.

Results-Oriented Approaches

Dixon (1987) stated that evaluation is conducted not only to get feedback which can be used to improve the course, but to justify the class to decision makers.
Mealiea and Duffy (1980) pointed out that trainers must keep management aware of the contribution training makes to bottom line results. If through lack of concrete information, managers come to see training as a luxury or a net drain on resources, training and trainers will be axed early on in hard times (Marshall, 1991; Birnbrauer, 1990; and Bushnell, 1990). Birnbrauer pointed out that despite the political necessity of training evaluation, fewer than half of all trainings are evaluated.

Dixon (1987) suggested collecting base-line productivity figures before training and comparing them with similar figures after training. She denied the need for control groups, but admitted the necessity of connecting increased productivity with the training itself. The mere fact that increased productivity occurred during and after training does not mean that the training itself is the cause. In an attempt to make the case unchallengeable, she suggested supplementing productivity information with proofs that the participants learned something and proofs that the new knowledge was applied to the job.

Salinger and Deming (1982) recommended a cost-benefit analysis. The trainer calculates the costs of not providing the training, then calculates the cost of providing the training. If the cost of continuing without training is greater than the cost of the training itself,
then the training is cost-effective. As Mealiea and Duffy (1980), however, pointed out that cost-benefit approaches only work well at low levels in which numbers can easily be paired with productivity. Doing a cost-benefit analysis on managers can be extremely difficult.

Stein (1981) described cost-effective programs as the product of a complete and realistic assessment of the gap between present performance and desired performance, based on prescriptive rather than felt learning needs. Stein used the term prescriptive as a synonym for objective. By felt learning needs, he meant learning needs based on an interest in the subject rather than objective necessity.

Cascio and Morris (1990) argued that personnel staff must think like business people since they are part of a business and must communicate with business managers. They maintained that training evaluators should use techniques employed in financial circles. The specific technique they favor is Net Present Value (NPV), since it is commonly used in capital budgeting to evaluate competing investment proposals. For a training to be justifiable, NPV must be positive. The calculations must also use discounting to account for the effect of the decaying benefit of training over time.

A decision to develop results-oriented evaluation was viewed with doubt by McEvoy and Buller (1990). They related how their own calculations using Cascio's (cited
in McEvoy & Buller, 1990) utility formula led to such "whopping" results as to be highly questionable by strict accounting-oriented clients. They doubted the believability of the claims their calculations seemed to extol for their particular training format and procedures. **Quality Control Approach**

Private industry now spends as much on education and training as the entire public school system, according to Carnevale (cited in Brinkerhoff, 1989) and Eurich (cited in Brinkerhoff, 1989). Brinkerhoff (1989) expressed concern, however, that training as a profession would not survive unless training departments learn to apply quality measures to the total process of training evaluation. He challenged training professionals to make training a valuable partner to the business of organizations, whether the business be producing products that make a profit for shareholders, delivering a quality education, or making appropriate placements in an adoption agency.

Several trends which impact training perspectives were listed by Brinkerhoff (1989). Transition to a global market, new manufacturing technologies (e.g., just-in-time inventory), and computer assisted design all contribute to extreme competition and increasingly rapid product development cycles. Work forces must be adaptable and proficient in new technologies. Brinkerhoff stated a belief that effective managers must become change agents,
capable of leading rather than simply adapting to change. He cited the push for customer satisfaction and service quality as reasons that industry is propelled forward by employees with certain skills and attitudes.

Pervasive administration of trainee satisfaction measures is useful in an immediate sense, i.e., to gauge trainee interests and to gather their suggestions, stated Brinkerhoff (1989). Long term, though, he felt that depending on trainee satisfaction indicated a lack of understanding of the real customer of training services. The true customers of training are upper-level managers. Their satisfaction with training stems from the extent to which training contributes to profit and to gaining a competitive edge. If measurement drives and reflects performance, then what gets counted, according to Brinkerhoff, must count.

In alignment with the emphasis on customer service, Brinkerhoff (1989) recommended treating top management personnel, whom he identified as the customers of training, just as other customers are treated. Suppliers of training should keep in contact with the customers, understand their needs and expectations. Just as with other customers, surveys of customer satisfaction, studies of customer needs and priorities, research into customers' problems and issues of importance are essential tools of evaluation. On the other side of the coin, training-
evaluation reports to management couched in terms familiar and important to them, can gain needed understanding and support for the training program. The ultimate effect could even be management sharing responsibility for training effectiveness.

Advising that training departments use quality control and management approaches, Brinkerhoff (1989) recommended analyzing and identifying quality criteria for each step of training. He also urged creation of a developmental environment. In such an environment, training personnel would treat every training program as a pilot program, subject to amendment as the environment and needs changed. Localized variations would make training modules even more adaptable. A final admonition was to demonstrate a commitment to training alternatives. By so doing, trainers would consider the possibility of making changes in work processes, implementing efficiencies, altering policies, etc, in cases for which training was not the best solution.

A continuous improvement system for training and development was also described by Dixon (1992). The system was built upon jointly establishing customer requirements, agreeing upon how to measure those requirements, collecting the measurement data, and making appropriate changes based on the results. Evaluation was based upon a map which visually displays the relationship
between training, as delivered by the training function, and business objectives, determined by management and shared with the developers of training.

The Importance of Needs Assessment

A shift to needs-driven versus program-driven training was among changes which Brinkerhoff (1989) recommended. Program-driven training simply paralleled whatever programs were in force at any given time. He defined needs-driven training as training based on an adequate needs assessment and designed to fill gaps in knowledge, skills, or abilities needed to meet organizational goals. The following evaluation questions were hailed as crucial: (1) What training goals, current and future, will best serve business needs and strategic priorities? (2) To what extent are current training results sufficient to support business needs? (3) What additional training would best serve business needs? (4) What revisions and improvements to existing training programs (including deletions) would best serve business needs?" Questions such as the above clearly indicate that trainers need to be well-informed regarding corporate strategy, problems, and priorities. Correctly or incorrectly, Brinkerhoff also argued that needs-driven training would reduce training inventory.

The basic categories of needs assessment discussed by Birnbrauer (1990) were needs identification forecasting
and needs analysis. In needs identification forecasting, the trainer gathers data from statistical measures of productivity. Whatever the company is producing should be measurable and, if it is measurable, management should have measured it. The trainer can also gather performance and behavioral indicators relating to the workers. All the items Birnbrauer listed in this category were statistics, such as those relating to absenteeism, tardiness, injuries, promotions, pay increases, requests for transfer, turnover rate (attrition), and performance appraisals. Finally, he mentioned safety statistics. Safety statistics include accident and injury rates, statistics regarding safety violations, and interventions by OSHA, all of which are objective indicators.

While the items gathered in needs identification forecasting are objective, Birnbrauer (1990) implied that needs analysis gathers subjective indicators. He stated that the methods of gathering information for needs analysis are: (1) questionnaires and checklists, (2) individual and group interviews, (3) observation, including critical incident observation, and (4) work participation, which he recommended only for entry level or simple tasks.

In their instructional technology model, Goldstein and Gilliam (1990) defined needs assessment as the first component. The needs assessment is used to design the
training program and the evaluation model. In their analysis, needs assessment consists of a series of five steps. The five steps are: (1) organizational analysis, (2) task and KSA analysis, (3) person analysis, (4) training environment design, and (5) evaluation.

Trainers do an organizational analysis by examining the organization's goals, the climate for training, and constraints present in the organizational environment which might prevent training from being successful, according to Goldstein and Gilliam (1990). The organization's goals are set by top management. They are frequently written down. If not, the trainer can elicit them from the responsible managers. Constraints often relate to politics. Supervisors might oppose training or the job performance concepts taught in training, for example. Constraints can be resource issues also. For example, training people to operate a machine that is not available is useless. Resource constraints are probably ultimately political as well. An illustration might be: if a machine is not available, can be shown to benefit the company's earnings, and the decision to purchase it or not will be a political decision.

Task and KSA analysis is an analysis of what should be taught and when in the process. The third step, the person analysis, evaluates the skills and capabilities of those individuals who are to be trained. The trainers use
the results of this analysis to focus training on critical KSAs. The fourth step is training environment design. Finally, the fifth step in needs assessment is evaluation. Participants may perform positively in training, and yet their positive training performance may not correlate with improved job performance. For this reason Goldstein and Gilliam (1990) think of training as a closed-loop feedback system, in which the result of the training is evaluated and the results used as feedback which may result in modifying the training itself.

As Mealiea and Duffy (1980) pointed out, the needs assessment may unearth reasons to believe that training is not needed. Their position was that managers call for training after observing a performance deficiency. However, performance deficiency is a symptom, and training is only a solution if skill deficiency is the underlying problem. If the deficient employees have the skills they need, the authors fault the incentive system, which they believe has one or more of four causes: (1) rewards for poor performance, (2) sanctions against good performance, (3) rewards and sanctions not based on performance (in which case it seems a deficiency would not be noticed), and (4) performance hampered by circumstances. According to Mager and Pipe (cited by Birnbrauer, 1990), if the employee could not do the job at gunpoint, he or she needs training. Otherwise, the problem is with incentives.
In a review of current literature on training, Tannenbaum and Yukl (1992) reported that the importance of conducting a thorough needs analysis was well accepted in the literature. While most organizations support the idea of needs analysis contributing to instructional objectives and training criteria, Tannenbaum and Yukl found that only 27% of companies surveyed by Saari et al in 1988 reported having procedures for determining the training and educational needs of their managers. They referred those interested in a thorough description of training needs-assessment methodology to Goldstein et al. (1991) and to McGehee and Thayer's (1961) framework of organizational, task, and person analysis categorization.

The original purpose of organizational analysis, according to Tannenbaum and Yukl (1992) who cited McGehee and Thayer (1961), was too provide information about where and when training was needed in an organization. They referenced Goldstein to explain that, reconceptualized, organizational diagnosis now examines system-wide components that determine if training can yield behavioral changes back on the job. Tannenbaum and Yukl reported a recurring theme of need to liken training and organizational strategy, just as did Sonnenfeld and Peiperl (1988), and Shuler and Jackson (cited in Tannenbaum, 1992).
A needs assessment strategy recommended by Kaufman (1983) purports to show that needs assessment does not cost; it pays. Having decided to plan using data from a needs assessment, the training evaluator is encouraged to select a needs assessment and planning level, either middle, comprehensive, or strategic. The middle level hinges upon concern for organizational resources (or inputs), plus the procedures and methods (or processes) to be employed in organizational activities, plus the immediate results (or products) accomplished.

The comprehensive level explained by Kaufman (1983) combines the middle level concerns with the products or services (or outputs) an organization does or does not deliver to its clients. It includes what an organization uses, does, and delivers to itself as well as to its external clients.

The strategic level combines the concerns of the comprehensive level with a consideration of how useful the organization's contributions (or outcomes) are to its clients and to the world in which its contributions must function. What the organization does and delivers as well as the impact these have on clients and on society in general.

Identifying needs assessment and planning partners follows selection of a planning level, according to Kaufman (1983). Groups who must be included for buy-in
are those who will be affected by the results, those who will implement the plan, and clients or society that will receive the results. Planners may need to do stratified random samples of the three groups, and should avoid loading with friends or tokens.

"Need," in Kaufman's (1983) needs assessment, refers to a gap in results. "Quasi-need" refers to a gap in processes, methods, inputs, or resources. Humans supply judgments concerning perceived needs (soft data). Hard data is performance-based (measurable). Examples include: profits, organizational image, death rates, numbers of people with positive credit ratings, returns to healthy functioning, quality of community life, and the like, all external outcomes. Organizational performance indicators such as productivity, rejection rates, case completions, absenteeism, morale, and complaints are internal sources of hard data.

While extolling the need for gaining participation of the planning partners, Kaufman (1983) gave some suggestions. When contacting the partners tell each of the expectations, time commitments, desired products, and required contributions. Reveal how much you'll support them—with funds, travel, data, materials, secretarial services, and the like—and how you'll use and consider their inputs. First meetings can be face-to-face, written surveys, delphi techniques, teleconferencing, or computer
interface. Inactive partners should be replaced with active participants.

Trainers must obtain acceptance of the needs assessment and planning level by explaining the levels, Kaufman (1983) emphasized. Needs should be explained as gaps between what is and what should be, whether relating to inputs, processes, products, outputs, or outcomes. He pointed out the need for reliability and validity in collecting both internal and external needs data. The two sources recommended were the perceptions of the planning partners and the actual performance discrepancies collected from controlled, objective observations. Tools recommended for sensing were: rating scales, questionnaires, structured interviews, or other paper-and-pencil assessments.

In addition to people's perceptions, Kaufman (1983) suggested collection of hard data such as: (1) absenteeism, (2) production rates, (3) cases closed, (4) audit exceptions, (5) courses completed, (6) certified competencies, (7) sick leave, (8) work samples, and (9) rejection rates. If the strategic level of needs assessment was chosen, the following internal and external data might be considered: (1) clients' perceptions and satisfaction, (2) return rates, (3) recidivism, (4) complaints, (5) profits, (6) return on investment, and (7) income.
Agreement is needed, explained Kaufman (1983), to list identified and documented needs. He stated that is helpful to ask what the situation would look like if the perceived gap were closed. He warned to discourage solutions being pushed by special groups and to encourage all partners to define needs as gaps in results. The end product would be an actual list of gaps in results.

Planners must place needs in priority order and obtain agreement of partners, according to Kaufman's (1983) final directive. He suggested that criteria for weighting might actually be dollar amounts attached to the costs of meeting the need and the cost of not meeting the needs.

Mapping can provide an effective needs assessment process, stressed Dixon (1992). The map serves an important role in gaining agreement from relevant stakeholders on the information in each column. Management strategy is clarified in the process. Dixon emphasized that training personnel would need to be persistent to facilitate definition of true behaviors. When managers propose as examples of needed behaviors "strong product knowledge," the facilitator would need to ask, "What do employees (sales representative, engineering technicians, etc.) do when they have strong product knowledge? How do they function differently from those who do not have strong product knowledge? If I observed
them, what would they be doing?" According to Dixon, management ownership of all items supports training efforts toward identified needs and behaviors which satisfy those needs.

In the present study, examples of needs assessment activities were observed to be strong at the actual production level. Supervisors were included in course design and refinement. As the study progressed, a needs assessment of higher management in the production area was instituted. It is the belief of the present researcher that earlier involvement of top manufacturing management would have greatly facilitated research efforts and would have possibly given training evaluation efforts a different focus.

Levels of Evaluation

Several methods of classifying training evaluation were discussed by MacKinney (1957). In order of ascending effectiveness, he listed evaluation of: (1) the trainee’s classroom behavior, (2) the trainee’s on-the-job behavior, and (3) on-the-job behavior of the trainee’s subordinate (Lindbom & Osterberg, cited in MacKinney, 1957). MacKinney also cited a simple pretest and posttest design as being a low level, with addition of a pretested and posttested control group forming the ultimate in experimental design.
A three-unit scale was discussed by Goodacre (cited in MacKinney, 1957). The lowest level was trainee attitude. The second level was knowledge acquired by the trainee. Goodacre listed evaluation of actual job performance as the highest level.

In regard to criteria of evaluation, MacKinney (1957) labeled objective performance scores (e.g., output automatically measured by a machine) as the most relevant. Objective performance, however, is difficult to obtain without the danger of confounding by variables other than training contributing to measures. Subjective evaluation was rated as less pertinent, but capable of being done in a thorough and systematic manner.

Kirkpatrick’s Four Levels of Evaluation

Birnbrauer (1990) stated that the most commonly used model for evaluation is the Kirkpatrick model. Kirkpatrick listed four levels of evaluation which are appropriate to training in business settings: reaction, learning, behavior, and results.

Level One Evaluation: Reaction

Birnbrauer (1987) divided measures into two categories. Criterion-referenced measures measure performance of trainees or programs according to performance criteria which are preset prior to measurement. Norm-referenced measures compare individual trainees, groups, or programs to other individual
trainees, groups, or programs. Level One evaluation is usually done with a criterion-referenced measure.

Kirkpatrick (1959, November) defined the reaction level of evaluation as measuring how well trainees liked the training. Reactions to subject content appropriateness, style of presentation, and physical arrangements were typical items. Although relatively simple to obtain and almost universally administered, according to Kirkpatrick (1959, November), the reaction level of evaluation should be safeguarded by several precautions. In addition to very obvious requirements, Kirkpatrick recommended that reaction evaluation forms be: (1) easily tabulated and quantified, (2) anonymous in order to assure honestly, and (3) constructed in a manner which allows conferees to write additional comments.

Kirkpatrick recommended a procedure seldom utilized by most trainers. He suggested administering reaction forms early in the training format, especially for long trainings. If, for example, reactions obtained after a third session reflected some sort of disapproval, changes could be made in the remaining six sessions.

Reactions to training, according to Kirkpatrick (November 1959), can be considered important beyond simply making the trainee and trainer feel good. He stated that a learner who likes the training will learn more. He also pointed out that the opinion and influence of a few key
people may determine whether or not the training will be continued.

The reaction form, which Alliger and Horowitz (1989) call a happiness sheet and Birnbrauer (1990) calls a smile test, is a questionnaire which elicits whether the trainee had a good time during the training. Many trainers use reaction forms on the theory that anything is better than nothing—a sort of evaluation by default. But Dixon (1987) and Alliger and Horowitz (1989) disagree. Dixon stated that it would be better not to evaluate at all than to use participant reaction forms. Birnbrauer (1990) had a more positive attitude, reasoning that learning is facilitated by a pleasant attitude toward the training.

Alliger and Horowitz (1989) tested the value of reaction forms by correlating reaction forms with the results of a gain-score test for a training. The gain-score method simply compares pretest and posttest data and measures gain in relevant KSAs (knowledge, skills, abilities). They found that correlations between reaction forms and gain-scores were very low, indicating that reaction forms may not be valid indicators.

Dixon (1987) identified three problems with reaction forms, and, by extension, with Level One evaluations. The first problem is that persons who control the trainer's rewards will make those rewards contingent on his or her ability to get positive responses on reaction forms. The
reward in question is continuing employment, whether the trainer is a contractor or an employee. Reaction forms thereby make a trainer's continuing employment indirectly contingent on his or her ability to put on an entertaining show and call that "training."

Dixon (1987) argued that by reinforcing entertainment rather than learning, smile sheets cause trainers to concentrate on showmanship rather than instruction. This criticism is soundly based in learning theory. However, she also takes the more questionable position that entertainment is "often" mutually exclusive with learning. She is careful to use the equivocal word "often" rather than an unequivocal word such as "always," but part of her argument is that smiling participants "often" have not been instructed. She contends that learning is "often" confusing, uncomfortable, and stressful, and she suggests that it "often" threatens the identity of the participant, especially if the seminar has to do with management development.

The second and third problems Dixon (1987) identified with regard to reaction forms are closely related. The second is that they point to the effort on the part of the trainer and course designer, and not to the effort made by the participant. The third is that they communicate the idea that the active involvement of the instructor is important, and that the active involvement of the
participant is not important. Consistent with her belief that Level One evaluations are counter-productive and should be discarded, she prescribes several Level Two evaluation techniques as ways to solve these problems.

**Level Two Evaluation: Learning**

Kirkpatrick (December 1959) defined the learning level of evaluation as measuring whether principles, facts, and techniques were understood and absorbed by trainees. His guideposts pointed out the need for quantitative, objective, before-and-after measures in cases for which facts and principles were the targeted learnings. He also recommended the use of a control group and statistical analysis to assure that learning was derived from training.

In the case of evaluating skills learned, Kirkpatrick (December 1959) suggested that training directors devise their own methods. He explained that many demonstrations, presentations, and exercises could be used to ascertain the degree to which skills were learned.

For Level Two evaluation Salinger and Deming (1982) recommended the delayed treatment design. The delayed treatment design is designed to neutralize the so-called halo effect, or rating one aspect favorably because of liking for another aspect.

Salinger and Deming (1982) described the halo effect as the effect on a manager's judgement of a prior belief
or experience about a person regarding his fitness. Highly subjective judgements are particularly susceptible to the halo effect. Bernardin and Beatty (1984) claimed that behaviorally anchored rating scale (BARS) can at least inhibit the halo effect if the behavioral descriptions attached to the BARS are as unambiguous as possible. In arguing for delayed treatment Salinger and Deming (1982) disputed effectiveness of BARS in such cases. In their estimation, judgements regarding management skill and ability remain highly subjective even when BARS are used.

The "delayed treatment" design Salinger and Deming (1982) described uses a randomly selected control group which is required to wait out the first cycle of the training program. After the first cycle is completed, both groups are assessed. Using a control group washes the halo effect out.

After the first training cycle, Salinger and Deming (1982) stated that the control group should be trained, and at the end of the second training cycle only the control group should be assessed. This makes it possible to evaluate changes in the training cycle as a result of the first assessment.

Dixon (1987) suggested pre- and post- measures of learning which measure performance and not course content. An example she cites is videotapes used in effective
presentation seminars. In addition to evaluating the training itself, pre-testing increases the amount of learning in a course, while post-testing causes participants to retain information longer.

Dixon (1987) also suggested performance measures, which include role playing, cases, and problem solving. If performance measures were not appropriate or effective, she suggested tests. Both tests and performance measures should be designed to give the instructor feedback on whether objectives are being met.

A third evaluation technique Dixon (1987) suggested was group interviews. Thirty minutes before the end of the course, ask participants to say what learning they will be able to put to use immediately. The group interview is different from the smile sheet. It emphasizes utility, not entertainment, and avoids use of numbers, which means the instructor is not evaluated in such a way that rewards can be bestowed or withheld. She also suggested follow-up group interviews after the training is completed to reinforce training and allow group support and learning.

Level Three Evaluation: Behavior

Kirkpatrick's (January 1960) third level of training evaluation was described as measurement of the degree to which trainees put the training into practice back on the job. The focus of measurement was actual job behaviors.
As Kirkpatrick explained, evaluation of job behaviors is more difficult than reaction and learning evaluation. A more scientific approach is needed, and many factors must be considered.

To evaluate training programs in terms of behavioral changes on the job, Kirkpatrick (January 1960) advocated a systematic appraisal of on-the-job performance both before and after training. He stated that the appraisal should be made by one or more of the following groups: (1) the person receiving the training, (2) the person’s supervisor, (3) the person’s subordinates, and (4) the person’s peers or other people thoroughly familiar with his or her performance. Again, Kirkpatrick recommended a control group and statistical analysis to assure that behavioral changes were related to training.

Kirkpatrick (January 1960) cited methods which he considered effective from several studies. Most of the studies used simple experimental designs and incorporated questionnaires or opinion surveys. In one study (Stroud, cited in Kirkpatrick, 1960), respondents were asked for critical incidents in categories. In two of the studies cited (Moon and Hariton, and Buchanan and Brunstetter), respondents were asked for then-and-now (or before-and-after) responses as a post-test.

Salinger and Deming (1982) stated the purpose of a Level Three evaluation in the form of a question: To what
degree is learning transferred to the job? They then proposed three methods for answering this question: the modified critical method, over-the-shoulder evaluations, and performance analysis.

The modified critical incident method, according to Salinger and Deming (1982) was developed by Professor James N. Mosel of George Washington University. Using Mosel's approach, the trainers ask participants by letter to participate in an after-training evaluation one to six months after training ends. This part has its critics. Leifer & Newstrom (1980) fault delayed evaluations because too much time elapses before follow-up, especially if the new skills are not being used.

Trainers using Mosel's approach follow up the letter with a telephone call or visit, setting an appointment with each participant for a 30-minute to 60-minute interview (Salinger & Deming, 1982). At the interview, trainers ask standard questions, followed by "probes" (meaning probing questions). Trainers ask for behavioral changes that have taken place as a result of the training. Since the participants are self-reporting, trainers should ask for a specific historical example for each perceived behavioral change. The specific historical examples are called critical incidents. Trainers should also ask about obstacles to change and the impact of changes on the working environment.
In Mosel's (cited in Salinger & Deming, 1982) method, trainers write each behavioral change cited on index cards. They then sort the recorded behaviors by course module, related instructional objective, or by some natural grouping which can be named. Then they determine behavioral depth and breadth. Depth is defined as number of behavioral changes class-wide. Breadth is defined as number of participants who changed. Optionally, trainers may list the behaviors cited in each category. Finally, they should describe other change phenomena, such as obstacles to change, impact of change on the working environment, and non-behavioral changes, such as knowledge, attitude, and planned future changes. The modified critical method has three advantages: (1) no pre-course evaluation is required, (2) it takes less time than a method which does require pre-course evaluation, and (3) it is very versatile.

The second strategy Salinger and Deming (1982) recommended for answering the Level Three question was "over-the-shoulder" evaluation. They defined over-the-shoulder evaluation to mean evaluating similar or identical trainings given less than six months earlier. Supposedly, evaluating a previous training of identical or similar content would yield quicker information than waiting to evaluate the focus training. Over-the-shoulder evaluation is undesirable to do on courses taught more
than six months earlier because the data may be corrupted by human memory lapses and outside influences. Two trainings must not be assumed to be precisely comparable just because they have the same subject matter. The instructor may be different in each case, the course content and materials may be different, and significant differences may exist in the backgrounds of the participants.

Finally, the third strategy Salinger and Deming (1982) recommended for answering the Level Three question is performance analysis. What the authors call performance analysis is actually non-performance analysis, or analyzing the reasons for employee non-performance. Typical problems that turn up in a performance analysis are unqualified or untrained personnel, missing or non-functional tools, and a suboptimal or counter-productive incentive system. The trainer's goal in doing a performance analysis is to determine the strategies that are required for correcting non-performance, and to determine whether or not training is one of those strategies. The trainer decides whether or not to do a performance analysis based on such considerations as timing (politics), the relative importance of the problem, and the available resources.

Salinger and Deming (1982) also posited another question which is a part of a Level Three evaluation: To
what degree is the acquired knowledge or skill maintained over time? For measuring, they recommended time-series evaluation, wherein multiple samples of completed work are taken before and after training. By comparing and contrasting samples, the trainer determines whether skills which result in improved performance have been transferred. Degree of performance deterioration over time indicates mastery or lack of mastery. Time series evaluation is primarily useful for repetitive tasks which are easily quantified.

Dixon (1987) suggested that trainers choose Level Three Evaluation strategy according to how frequently the skills to be measured are used on the job. If the skills to be measured are used all the time, she suggested surveys. If they are used only at specific times or in specific settings, she suggested observation. If they are only used infrequently or spontaneously she suggested the critical incidence method.

Level Four Evaluation: Results

The fourth level of training evaluation in the model of Kirkpatrick (1960) was the results level. The objectives of most training programs, he stated, can be identified as results in which the total organization is interested. Such objectives as reduction of costs, reduction of turnover and absenteeism, and increase in quality and quantity of production are examples of results-oriented objectives.
Kirkpatrick included concepts such as morale because of an assumption that morale will lead to organizational results. Kirkpatrick (1960) cited several studies involving level four evaluation of training. A training targeting accident rate reduction was credited directly with reducing accidents because no physical changes had been instituted to reduce accidents. Although a few other successful studies were cited, Kirkpatrick concluded that level four evaluation lacked adequate techniques to separate training from other causes in large-scale organizational changes. He called for training directors to develop level four evaluation techniques and to share them openly through journal contributions.

Level Four evaluation determines whether the training produced results which support the goals of the organization itself. Organization level goals include items such as lowered absenteeism. At a higher level, they also include increased profit for the company. Birnbrauer (1990) advocated what he called a systems approach, looking at an organization as a hierarchical collection of systems, subsystems and supersystems which are functionally related. Training is a system within the organization which supports other systems. By contributing to the goals of the departments it serves (viewed as systems) it contributes to the goals of the organization itself.
Level Four evaluation also includes cost-benefit analysis, which was discussed earlier. A high level organizational goal is to maximize profits. If training produces more measurable benefit (quantified in dollars) than it costs, then it maximizes profits, and promotes organizational goals.

Critique of Kirkpatrick's Model

Kirkpatrick's model of training evaluation served a valuable function by giving trainers a vocabulary and a taxonomy of training criteria. Alliger and Janak (1989) commended the simplicity of Kirkpatrick's model and its ability to help people think about training evaluation criteria. They acknowledged that the model served a felt organizational need. However, they questioned the validity of the three assumptions which seem to underlie the model.

An implicit assumption of Kirkpatrick's hierarchical model is that each succeeding level is more informative than the last. This assumption was examined by Alliger and Janak (1989). Alliger and Janak acknowledged that many well-known experts champion the dollar estimation of intervention effectiveness as the ultimate form of evaluation. They also acknowledged that trainers may feel a need to justify their existence in terms of contribution to organizational goals. They gave an example of at least one case, however, in which a plant manager opted against quantitative measurement in favor of highlighting the intervention itself. They also
cited Goldstein as stating that purely quantitative measures may obscure the range of available approaches.

The second assumption of Kirkpatrick’s model, according to Alliger and Janak (1989), is that each level is caused by the previous level. Countering this assumption, they cited Lewicki’s studies which indicated that attending can be deleterious to learning. Alliger and Janak cited research in educational classrooms (Remmers, Martin, & Elliott; Rodin & Rodin, cited in Alliger & Janak, 1989) in which Levels 1 and 2 were negatively correlated and in which humorous lectures are liked better (Level 1) but do not cause more learning (Level 2) (Kaplan & Pascoe, cited in Alliger & Janak, 1989). Alliger and Janak agreed that causal links do exist between learning, behavior, and results. They argued, however, that while behavior may lead to results, results in the form of feedback, may also sustain behavior, thus diminishing somewhat a hierarchical perspective.

The third assumption of Kirkpatrick’s model which Alliger and Janak (1989) examined was that each succeeding level is correlated with the previous level, or more generally, that all correlations among levels are positive. To refute this assumption, Alliger and Janak reviewed 214 articles on training evaluation. Of the 32 which reported on actual studies, only eight reported any correlation between two or more levels. Following more leads and references, Alliger and Janak finally found 12 articles
reporting significant correlational relationships between
two or more levels in industrial settings, still a fairly
limited number.

Although according recognition to Kirkpatrick's model
as a first, global heuristic for training evaluation,
Alliger and Janak (1989) suggested that growing
sophistication in the field may allow consideration of other
perspectives. One example they mentioned was the
possibility of utilizing other measures at the recognized
levels. One of the "other measures" they suggested was
evaluation by self, superiors, and subordinates. One could
be confused by such a suggestion because Kirkpatrick (1960)
recommended the same thing. Consequently, while expressing
a feeling that something is lacking in Kirkpatrick's model,
Alliger and Janak did not recommend any real improvements to
the model.

Issues in Performance Rating

Evaluation of training which teaches factual
information, such as the Physics/Chemistry course presently
under scrutiny, offers a challenge. How is one to evaluate
behaviorally that which is happening in a person's thinking?
Possible answers are offered by recent examination of the
cognitive processing and learning requirements necessary to
perform job tasks. Howell and Cooke (cited in Tannenbaum
and Yukl, 1992) recognized that task analysis in modern
corporations may require finding a way to analyze such
cognitive tasks as inference, diagnosis, judgment, and decision-making. The "think-aloud" verbal protocols and psychometric scaling methods of Cooke and Schvaneveldt 1988 (cited by Tannenbaum & Yukl, 1992) are the most common methods used in cognitive psychology. "Think-aloud" protocols may provide an excellent method to discover some of the ways in which training of factual principles transfers back to the job in the form of cognitive behavior.

A telephone conversation with D. L. Kirkpatrick (personal communication, September, 1992), father of the well-known four levels of evaluation, while offering no definitive step-by-step procedures, did provide a general framework for thinking about transfer of technical training. His comments follow:

What do you want them to do differently? Measure before and after. Interview the people themselves. Ask, "What are you doing differently?" Ask their bosses, "What are they doing differently?" It's like "leadership," "communications," "decision making" -- background for the way they behave. Impossible to absolutely prove, but measure before and measure after. Are we doing more of the things we think we should be doing?

Based on the common-sense idea of measuring before and measuring after, the present writer considered behaviorally anchored rating scales developed by asking trainees
themselves and their supervisors what the trainees were doing differently after training. Perhaps their own thinking aloud about behaviors before and after training would explain the difference training made on their job performance and provide anchors for evaluating performance.

**Behaviorally Anchored Rating Scales**

Cascio (1991) pointed out that narrative essays, employee comparisons, and graphic scales appeal to raters, but are not psychometrically acceptable. Forced-choice scales, conversely, are psychometrically acceptable, but do not appeal to raters. His solution was to strike a compromise between these two extremes. The compromise he proposed is behaviorally anchored ratings scales (BARS).

The BARS accomplishes all three of the criteria set by Birnbrauer (1987) for evaluation instruments. It puts meaningful labels on the data being measured. It spreads dimensional data in a meaningful way. And finally it rates the item being measured by levels or steps.

Evaluative rating scales anchored by examples of expected behavior were developed by Smith and Kendall (1963). Their approach was purported to establish reliable and valid rating systems which did not impose the values, interpretations, and beliefs of psychologists. They rather sought consensus from representatives of the rater group as to what examples of behaviors represented particular characteristics of performance.
Attempting to meet the needs of raters in extremely diverse situations and to supply face validity to the evaluation instrument, Smith and Kendall (1963) utilized a series of continuous graphic rating scales, arranged vertically. Behavioral descriptions, exemplifying various degrees of each dimension, were printed along a vertical line at different heights. Judgments of head nurses similar to those who would be expected to use the scales determined scale positions. The examples served as anchors to define levels of the characteristic, and as operational definitions of the dimension being rated. Raters were to check a position along the line and to support the placement with notes concerning actual observed behavior (Smith & Kendall, 1963).

The format chosen by Smith and Kendall (1963) was an attempt to combine the relevance of critical incidents and the acceptability to raters of graphic rating scales, while abnegating any use of forced choices as an alternative. However, raters were not asked to recall specific observations, but rather to predict future behaviors based upon past observations.

The use of expected behaviors was intended to encourage conscientiousness. Concrete predictions with previously agreed-upon definitions were intended to eliminate central tendency or hedging effects. Expected behaviors were made verifiable so that future divergent behaviors of ratees
would actually challenge the values, judgments, insights, etc. of the raters (Smith & Kendall, 1963).

Retranslation was a focal element of the BARS development procedures utilized by Smith and Kendall (1963). The process resembled situations wherein material is translated into a foreign language, and then, by an independent translator, retranslated into the original, with slippages demanding correction. The iterative work of one group was checked and revised by others within a restricted content area of medical-surgical nursing.

Initially, qualities or characteristics to be evaluated were listed by four groups of head nurses. The most frequent dimensions were selected for further analysis, retaining the nurses' own terminology. Gathering and categorizing critical incidents insured further coverage of pertinent dimensions. After formulating general statements representing definitions of high, low, and acceptable performance for each quality, the groups submitted examples of behavior in each quality. These examples were edited into the form of expectations of specific behavior (Smith & Kendall, 1963).

Independently, judges indicated what quality was illustrated by each example. Examples were eliminated if there was not clear modal agreement as to the quality to which each belonged. Qualities were eliminated if examples were not consistently reassigned to the quality for which
they were originally designed. Another group of judges used the examples to describe specific nurses with excellent and unsatisfactory performance. Differences between the pairs of nurses were computed to determine the discrimination value of the examples. As the scales were presented with the resulting definitions and items, still other raters assigned a value from .0 to 2.0 according to desirability. Items yielding too great dispersion or multimodal distribution were eliminated. Remaining items were assigned mean scale positions. Mean scale positions assigned by each group of judges were intercorrelated to determine scale reliability. The lowest scale reliability after elimination of scales with dispersed or multimodal elements was .972 (Smith & Kendall, 1963).

Cascio (1991) described BARS generation as a multi-step process. The first step is an initial conference. A group, which may be all supervisors or all workers, or a combination of both, identifies and defines all the important dimensions in effective job performance. A second group generates critical incidents for each dimension identified by the first group. They must come up with critical incidents for effective, mediocre, and ineffective performance.

The critical incident technique was developed by Flanagan (cited by Stein, 1981). He defined a critical incident as an "observable activity of sufficient duration
and complete in itself to permit inferences and predictions about the behavior of the person performing the act in a situation where the purpose or intent of the act is clear to the observer and where the consequences of the act are sufficiently definite to leave little doubt concerning its effects." The three components of a critical incident are an actor, an action committed by that actor, and a result which can be attributed to the action committed.

After the second group finishes the critical incidents, Cascio (1991) said a third group gets the output of the first two groups, but with the critical incidents randomized. Their task is similar to a matching test, in which they match critical incidents to dimensions and determine whether a particular critical incident indicates high, medium, or low performance. This is a quality check. It is inspired by retranslation, in which a work translated into a foreign language is translated by a second translator into its original language, and the two original language versions are compared.

According to Cascio (1991), if the judges in the third group cannot agree on the correct dimension for a critical incident, they must eliminate that critical incident. If they cannot assign any incidents to a particular dimension, they must eliminate that dimension. If they assign too many incidents to an unspecified "miscellaneous" dimension, they must create one or more new dimensions.
Lawshe (cited by Warrenfeltz, 1989) called the foregoing process content validation. To quantify the validity of content for each dimension, Lawshe suggested using the percentage of SMEs scoring an item either high or medium. He called the percentage, which is stated as a decimal fraction (i.e., .9 instead of 90%), a Content Validity Ratio, or CVR. The same author also suggested a Content Validity Index, or CVI, which is the average of all the CVRs.

After the third group is finished and content validation is completed, Cascio (1991) stated that a fourth group of judges should take the output and rate each critical incident on a scale of 1 to 9. Then the group calculates the mean and standard deviation for each dimension. Using the standard deviations, they quantify how much agreement or disagreement there is regarding the significance of a critical incident. If there is too much disagreement, the critical incident is discarded. This assures the unambiguous character of the ratings that Bernardin and Beatty (1984) believe is necessary for BARS to be useful. The need for lack of ambiguity is reflected in the name of the instrument itself. Unambiguous behaviors are said to be "anchored," rather than allowed to float according to the tastes and prejudices of the rater. Bernardin and Beatty also added that behaviors should be
selected which are spaced at equal intervals along the BARS scale.

After the fourth group is finished, Cascio (1991) said the fifth group takes over. This is a group of supervisors, who rate their subordinates according to the dimensions and critical incidents produced by the first four groups. Two or more supervisors rate each subordinate independently. They then compare scores for each subordinate, and cross-correlate scores for different dimensions to see if dimensions are independent or not.

Bernardin and Beatty (1984) showed the finished BARS as a vertical scale with ten levels of differentiation between maximum and minimum performance. The top of the scale represents top performance and the bottom of the scale represents low performance. Descriptions of various behaviors appear at various points on the scale to define what a particular number means.

Ash (1986) described two other instruments which are very similar to BARS. These are the Behavioral Consistency Application Supplement (BCAS) and the Activity/Achievement Indicator (A/AI).

The BCAS is based on Schmidt's (cited by Ash, 1986) observation that past behavior is a much better predictor of future behavior than past exposure. Reasoning from that premise, Schmidt argued that evaluators should find out what the applicant has done in the past, and not concentrate on
circumstances to which he or she has been exposed. The BCAS is a formal instrument for achieving this objective.

To construct a BCAS, Ash (1986) stated that Subject Matter Experts (SMEs) generate a statement of knowledge, skills, and attitudes (KSAs) for the job being evaluated, which they combine in five to seven achievement dimensions. The applicant then writes a description of at least two past achievements which demonstrate his or her capabilities for each of the achievement dimensions. Each description must include five items: (1) a statement of the problem or objective, (2) what the applicant did and when, (3) what the result was, (4) how much credit he or she takes for the outcome, expressed as a percentage, and (5) a name and telephone number for a witness who can verify the statements made.

SMEs then rate the achievements on quality and develop a scale similar to BARS. Training and work experience (T&E) evaluators then use these scales to evaluate applicants, and combine scores across dimensions to derive a single score for each applicant.

The BCAS has a problem, inasmuch as substantial numbers of applicants refuse to complete it. Moreover, research shows no correlation between refusal to complete the BCAS and applicant quality, but it does show that BCAS is a good indicator when it is completed. For BCAS to be valuable, it therefore must be used. To solve this problem, Ash
suggested a simpler system, on the theory that applicants will cooperate if evaluators use a less demanding measure. He called this less demanding measure the Activity/Achievement Indicator (A/AI).

The A/AI, according to Ash (1986), consists of sets of activity/achievement statements which are already written when the applicant comes to take the test. Each set represents a specific KSA dimension, and contains three statements, representing high, medium, and low achievement. These statements are quite similar to scale anchors for BARS, and are generated in the same way. The applicant selects one statement from each set which is most typical or his or her background.

The A/AI is therefore similar to BCAS but differs from it in three ways (Ash, 1986). First, A/AI does not require the applicant to write, so the results are independent of the applicant’s writing ability. Second, A/AI does not require that the applicant come up with critical incidents. It therefore attempts to measure typical past performance, whereas by eliciting critical incidents, BCAS attempts to measure peak past performance. Finally, A/AI requires the applicant to assess himself or herself.

The approach to behaviorally anchored rating scales (BARS) developed by Smith and Kendall (1963) advocates features not common to many approaches. In 1986, Smith and Kendall introduced a BARS model intended: (1) to encourage
raters to observe behavior more carefully, (2) to infer the meaning of that behavior, and (3) to record observed incidents on a continuum of effectiveness for specific dimensions over a period of time.

Raters using the Smith and Kendall approach (1963) were to: (1) observe incidents, (2) determine to which dimensions the incidents related, (3) scale and record incidents on a page at a point corresponding to predetermined examples of levels of effectiveness, and (4) to summarize observations and make ratings at the termination of the period. Recording of observations over a period of time was intended to allow accurate summary of performance for the period and sufficient clarity for discussion with the ratee.

The objective of Smith and Kendall (1963) was to standardize not only the rating process but also the observation process, thus enhancing the validity of future observations. Smith and Kendall's model aided the observation process by supplying examples of three levels of competence (critical incidents) and a definition of the dimension which subsumed the examples. They argued that to simply select one most typical incident at the end of a period negated the needed for a graphic format altogether. In answer to criticisms of great time involvement for the long-term rating process, Smith and Kendall maintained that
their model required no longer than summated scales containing a much greater number of items.

Behavioral expectation scales resemble behaviorally anchored rating scales closely in their development, with a difference existing in rating expected behaviors rather than observed behaviors. A study of the psychometric properties of behavioral expectation scales (BES) in comparison to summated scales was reported by Bernardin (1977). In this study, researchers developed three rating instruments. The initial, rigorous process used to develop the BES was used as a springboard for two summated scales, one which used the operational definitions of the behavioral expectations of the BES, and one which was derived from the components of the performance dimension definitions of the BES. A rigorous study of psychometric properties revealed no greater interrater reliability or discrimination among ratees. Nor could lower halo effect or leniency error be reported for the BES.

Bernardin (1977), nevertheless, suggested that two factors in the study support the efficacy of using the BES approach. First, the study showed lower halo and leniency effect than previously developed examples of the BES. Bernardin implied that summated scales which were developed as rigorously as the BES (as in his study where initial BES procedures were used), would naturally be more psychometrically impressive than many summated scales. He
also referred to studies in which development of the BES precipitated desirable spinoffs in work settings (Schneier, Beatty, and Beatty, cited in Bernardin, 1977).

Retrospective Pretests as a Method

Psychologists consulting in business regularly face the need to measure the effects of training and other interventions in a cost-effective manner. While behavioral evaluations are preferred, self-report measures are often the only feasible approach. Self-report instruments are typically administered before and after an intervention and at the same times to a control group of individuals who have not experienced the intervention (Howard & Dailey, 1979).

For pretest and posttest comparisons to have meaning, the two sets of scores must possess a common standard of measurement (Cronbach and Furby, cited in Howard & Dailey, 1979). If the standard of measurement were to change for any reason, (e.g., as a result of increased awareness due to training) so that subjects viewed their perceptions or performance prior to the intervention differently that they viewed those same perceptions or performance retrospectively after training, then pretest and posttest measures would have lost their common metric basis.

Researchers must consider the possible confounding of evaluation efforts in self-report measures if the internalized scale of measurement is distorted. A shift in the standard of measurement from pretest to posttest,
whether due to increased understanding of the variable being measured or due to other factors, would yield invalid interpretation of intervention effectiveness (Campbell & Stanley, 1963; Caporoso, & Neale & Leibert, cited in Howard & Dailey, 1979).

Evaluation of the relative merits of a retrospective pretest-posttest design to the traditional pretest-posttest design was the purpose of a study by Howard and Dailey (1979). Evidence of response shift was found when pretest and "then" scores were compared for students in an interviewing skills workshop produced by the Interviewing Institute of the University of Houston. "Then" scores were scores collected from students after a week-long training as estimates of their capabilities prior to training.

In Howard and Dailey's (1979) study, Then/Post and Pre/Post self-report change scores were correlated with Pre/Post change in judges' skill ratings and behavioral incidents. All comparisons favored the Then/Post approach, with 9 out of 13 differences between Then/Post and Pre/Post change scores reaching significance. In every case, Then/Post measures of effectiveness of training were less conservative than Pre\Post measures. One may interpret the latter finding to mean that subjects shifted perspective, and therefore responses, because of a post-training changed perception of capabilities prior to training.
In an experiment designed to test response shift in pretest-posttest evaluation of training, Sprangers and Hoogstraten (1989) utilized a Solomon four-group design (Campbell & Stanley, cited in Sprangers & Hoogstraten, 1989) to examine the extent to which administration of a self-report pretest affects the self-report posttest and retrospective pretest scores. Retrospective pretests and behavioral pretests were used to control for response shift bias, a change of the subject's internal standard. Sprangers and Hoogstraten (1989) hypothesized that a reason for response shift is the level of information subjects have at pretest regarding the dimension on which they are asked to self-report. They found that effective training resulted in response shift and that a behavioral pretest administered prior to the self-report pretest prevented a response shift from occurring. Presumably, the behavioral pretest allowed subjects to gain information about their level of function prior to the pretest. Without the behavioral pretest, then, a response shift would be expected. Accordingly, the subjects would be expected to demonstrate a different self-report on a retrospective pretest than on a conventional pretest.

The response shift study of Sprangers and Hoogstraten (1989) utilized three analyses of variance (ANOVAS) across the four conditions of the Solomon four-group design: one for self-report posttest scores, one for retrospective
pretest scores, and one for objective posttest scores to examine the main effects of Pretesting and Treatment and their interaction. They found that the pretest significantly affected the self-report posttest scores and retrospective pretest scores. Their study showed that the treatment had no effect and that there was no interaction between treatment and pretest. As would be expected with no treatment effect, pretest and retrospective pretest scores were not significantly different. In other words, there was no response shift.

The purpose of some experimental interventions, as Sprangers and Hoogstraten (1989) explained, is to elicit a change in subjects' understanding or awareness of target concepts. Such insights would almost necessarily change subjects' estimations of their levels of performance in relation to the concepts as well. A change in internal standards, also called a response shift, would then have occurred.

Response shifts which occur in pretest/posttest self-report measures of interventions, according to Sprangers and Hoogstraten (1989), introduce bias into otherwise well-planned experimental designs. The same bias confounds posttest comparisons between experimental and control groups.

In two experiments measuring pretesting and treatment effects, Sprangers and Hoogstraten (1989) illuminated the
subject of response shift. Their first experiment, using a classic Solomon Four-Group design, showed no treatment effect, and consequently, as expected, no response shift. However, an effect was found for pretesting, a result also worthy of consideration.

Four possible reasons for pretest biasing effects were summarized by Hoogstraten (cited in Sprangers & Hoogstraten, 1989): (1) The pretest may arouse the curiosity of subjects and consequently induce a motivational effect; (2) pretesting may draw subjects’ attention selectively toward certain aspects of the treatment; (3) the pretest may hint to the subjects at the purpose of the investigator and may facilitate or inhibit change, depending on the subjects’ willingness to comply with the experimental demands; and (4) the information contained in the pretest may bring about an advertising or priming effect, consequently facilitating change.

In the second experiment of Sprangers and Hoogstraten (1989), which utilized six conditions and three experimental groups, response shift occurred only in the case of the one experimental group which was not administered a behavioral pretest. Behavioral pretest seemed to inhibit response shift. Specifically, the no-behavioral-pretest group was the only group in which the retrospective pretest was significantly lower than the traditional pretest. It seems logical that both treatment (or training) and behavioral
pretests increased awareness of actual ability or perhaps brought about different expectations about appropriate levels of performance. Sprangers and Hoogstraten concluded that researchers may discard pretesting and consider the retrospective pretest a better alternative.

Experimental Design and Statistical Analysis

The Solomon four-group design represents the first specific procedure designed to consider external validity factors, according to Goldstein (1986). The pattern is Pretest-Treatment-Posttest for Group I, Pretest-No Treatment-Posttest for Group II, No Pretest-Treatment-Posttest for Group III, and No Pretest-No Treatment-Posttest for Group IV (see Appendix A). Random assignment of participants to the four groups makes it possible to examine the effects of pretesting. The fourth group, with no pretest and no treatment, also allows examination of the effects of maturation and history, potential threats to internal validity.

Prior to introduction of the Four-Group Design by Solomon (1949; Solomon & Lessac, 1968), only experimental and control groups were utilized. Solomon added the additional groups (treatment and control groups with no pretest) to account for contamination by pretesting.

While Bunker and Cohen (1977) predicted pretest sensitization to be especially likely in organizations which provide technical skills and abilities training to workers
already having had exposure to the concepts, they found no main effect for pretesting. In the case of semiconductor technicians who came to their positions with two years of technical training, pretest sensitization does seem likely in measures involving self-report ratings. While supervisors were asked to rate the technicians in the portion of the experiment which gathered sufficient data to be analyzed, technicians were also pretested for a phase of the experiment which did not elicit enough information to be included in the study. Consequently, pretest sensitization is expected.

An important concept pointed out by Bunker and Cohen (1977) was the likelihood of intervening variables bringing out higher order interactions. Many variables which are not being tested may intervene to alter findings. Such variables should, as much as possible, be included in the design, or at least in the statistical analysis as covariates. Intervening variables constitute one of the greatest threats to scientific rigor (Goldstein, 1986). In business settings, where access to pertinent information is often not available to the researcher, intervening variables may play havoc with significant response elicitation.

Goldstein (1986) stated that for designs involving naturally assembled groups inclusion of a control group is preferable to no control group. The more similar the groups and their scores on the pretest, the more effective the
control becomes in accounting for extraneous influences. Although internal validity factors like history, pretesting, maturation, and instrumentation are improved with a control group, this design is susceptible to interaction between selection factors and maturation, history, and testing. Investigators must be sensitive to potential differences between the groups and to differential treatment of the groups. Two advantages of this design may be: (a) that it is less disruptive and reduces experimentation effect (sometimes called the "Hawthorne Effect"), and (b) that in some settings a naturally occurring group affords a larger subject population.

Goldstein (1986) viewed content validity as a first criterion in training evaluation. To further safeguard validity, pretests and posttests should be added, if possible. Addition of control groups would increase control of validity factors. He recognized the constraints placed on evaluation by organizational and environmental factors, stating that models should permit the extraction of the greatest amount of information within the constraints of the environment. Researchers should not allow themselves to be frozen into inactivity by the looming possibility of threats to validity.

Investigators who are interested in evaluating the effectiveness of training and development must consider statistical treatment of the measurement. A discussion of
change measurement by Arvey and Cole (1989) assumes that data will be available from experimental and control groups who are randomly assigned. The assumption of random assignment, and thus equivalence of groups, is not always reasonable in business and manufacturing settings. However, an overview of design and statistical methods which yield various advantages and disadvantages in an ideal setting is pertinent.

Statistical methods reviewed by Arvey and Cole (1989) were related to specific experimental designs. To analyze results of a posttest only design, they recommended an independent $t$ test to test the hypothesis that the two groups differed on the posttest scores. They warned that a lack of statistical significance could reflect the possibility that the training had no effect, that sampling error reduced the difference between groups, or that a Type II error was made through lack of statistical power.

Gain-score designs, according to Arvey and Cole (1989), compute the difference between pretest and posttest scores. Again, they recommended a $t$ test, this time to analyze the difference between the gain-scores. Inclusion of the pretest helps to account for prior differences among subjects on the variable of interest, thus increasing ability to detect true differences.

The repeated-measures analysis of variance (ANOVA) was compared to gain-score analysis by Arvey and Cole (1989).
In the ANOVA, the between-subjects source of variance is represented by a group variable (experimental versus control), and the within-subjects source of variance is represented by time (i.e., the pretest and posttest). Arvey and Cole's assertion that comparison of between-subjects and within-subjects results is equivalent to comparison of gain-score results (when the within-subjects variable has only two levels) is corroborated by Huck and McLean (1975).

Analysis of variance, according to both sources, lacks precision by failing to account for the possibility that the pretest and posttest will not be perfectly correlated.

Use of the analysis of covariance (ANCOVA) was favored by both Arvey and Cole (1989) and Huck and McLean (1975) to increase power to detect change when pretest and posttest correlations are substantial. The ANCOVA approach incorporates and weights the pretest score as a covariate. The weight reflects the extent of correlation between pretest and posttest. The ANCOVA design treats the pretest as a covariate in a between-group design. When pretest-posttest correlations exceed $r=.5$, the power of the ANCOVA procedure to detect treatment effects exceeds that of the posttest only, gain-score, or analysis of variance techniques. While the loss of one degree of freedom inherent in the ANCOVA approach may concern some researchers, Cohen (1977) suggested that samples as large as fifteen or twenty should be minimally affected.
Addressing reliability, Arvey and Cole (1989) noted that a decrease in reliability (that is, by means of a more heterogeneous population creating more variability) increases power. Paradoxically, increase in reliability (by reduction of error variance) can also increase power. However, assuming that true variance remains constant, Arvey and Cole concluded that greater reliability is associated with greater power in the posttest, gain-score, and ANCOVA designs.

When the type of training precludes objective measures, as in the case of attitude change or translation of technical (knowledge) training to actual workplace behaviors, self-report necessarily becomes the measurement of choice. In self-report measures, problems arise in the definition of change (Arvey & Cole, 1989). According to Golembiewski, Billingsley, and Yeager (1979), three kinds of change occur with self-report measures.

The obvious change explained by Golembiewski et al. (1979) was a change due to training intervention. A comparison of pretest and posttest measures is typically undertaken to determine the change due to training. A more accurate measure, according to Terborg, Howard, and Maxwell (1980), is comparison of posttest to a retrospective pretest score (also called a then-score by Howard & Dailey, 1979).

Another type of change referred to by Terborg et al. (1980), was the type in which true change is confounded by
recalibration of the measurement scale. Howard and Dailey (1979) labeled the same phenomenon as response-shift bias. Response-shift bias occurs when a subject sees oneself as having improved after training, but then measures himself or herself against a higher standard. A confounding, or cancelling out effect, occurs. To detect a response-shift bias, Terborg et al. (1980) recommended a T score comparison of pretest and retrospective pretest.

Another change explained by Terborg et al. (1980) was reconceptualization of the construct of interest itself. Performance of tasks, or application of principles learned, may be represented differently than was the case before training. To capture the latter change, Terborg et al. suggested computing the correlation and change in variance between pretest and retrospective pretest items.

The present evaluation study is based on administration of behaviorally anchored rating scales (BARS) via the Solomon Four-Group Design. The BARS, based on a complex development process, is designed to capture four levels of performance on eight dimensions. Greater differentiation is attempted by spreading the four levels over a wider spectrum, yielding a ten-point scale. Retrospective pretests are administered to all groups in an attempt to determine the likelihood of a response shift bias. In the case of electronics technicians who came to their positions with two years of technical training, pretest sensitization
does seem likely in measures involving self-report ratings. However, based on the findings of Bunker and Cohen, pretest sensitization is not predicted, especially in light of being rated by their supervisors. Hypotheses follow:

**Hypothesis 1.** For technicians in semiconductor manufacturing companies, technicians receiving training in Physics/Chemistry classes demonstrate higher scores on the eight dimensions of the Behaviorally Anchored Rating Scales (BARS) for Semiconductor Technicians posttest, as rated by their supervisors, than technicians who do not receive the training. The eight dimensions of the BARS are: (1) Problem Solving, (2) Communication with Supervisor, (3) Communication as Part of a Group, (4) Written Communication, (5) Communication with Support Groups, (6) Troubleshooting, (7) Calculations involving Process, and (8) Understanding Total Process.

**Hypothesis 2.** Technicians whose supervisors rate them on a traditional pretest, score the same on the posttests as technicians whose supervisors do not rate them on a traditional pretest.

**Hypothesis 3.** Supervisors rate technicians differently on a retrospective pretest than on a traditional pretest, thus demonstrating a type of response shift bias.
CHAPTER II

METHOD

Subjects

Subjects were 103 engineering technicians employed by a semiconductor manufacturing company who were located at semiconductor wafer fabrication units in Dallas, Sherman, and Lubbock, Texas. The technicians were employed either in processes necessary for the manufacture of products or in operating and maintaining manufacturing equipment. Subjects were chosen on the basis of eligibility for and appropriateness of receiving training in Physics/Chemistry classes offered by the organization's training and development department. No discrimination was made between process engineering technicians and equipment engineering technicians.

Demographic Variables

Data on demographic variables were obtained through the process of collecting technical information through personal and telephone interviews. Worksites for the technicians in the study were three sites in Dallas (with 28, 21, and 14 technicians at sites 1, 2, and 3 respectively), one site in Lubbock (with 22 technicians), and a site in Sherman (with 18 technicians). Technical working conditions were similar.
in all sites. Educational levels were comparable, since all technicians were required to have a two-year associate degree in electronics as a condition of employment in the position. Of the 103 technicians in the study, 90 were males and 13 were females.

Materials

Materials consisted of: (a) behaviorally anchored rating scales developed for the evaluation project, (b) Scantron forms and Scantron software and hardware capable of reading the forms, and (c) an extensive internal corporate communication system. The behaviorally anchored rating scales were developed via procedures borrowed largely from Cascio (1991), with some adaptations to meet localized contingencies.

Dependent Variables

The eight dimensions on the Behaviorally Anchored Rating Scales for Semiconductor Technicians comprised the variables expected to be altered by training in Physics/Chemistry classes. The dimensions included: (a) Problem Solving, (b) Communication with Supervisor, (c) Communication as Part of a Group, (d) Written Communication, (e) Communication with Support Groups, (f) Troubleshooting, (g) Calculations involving Process, and (h) Understanding Total Process.
Independent Variables

Independent variables for the present study included: (a) training in Physics/Chemistry, with the two levels of training or no training, and (b) traditional pretest, with the two levels of pretest or no pretest.

Procedure

The researcher received a commission from the national director of training to evaluate technical training in terms of carryover to work behaviors. The evaluation called for an appraisal to determine the level of transfer from classroom training to the job, which is considered a level-three evaluation. Final products of the project were defined to be: (a) a valid evaluation instrument which could either be used for a number of classes or which could serve as a prototype for development of similar instruments, (b) a process for development and administration of the evaluation instrument, and (c) the data obtained through the process of administering the instrument.

By agreement with training managers responsible for developing and delivering technical training, Physics/Chemistry classes were chosen as the focus of evaluation. The choice was based on the likelihood of Physics/Chemistry being the most difficult training to relate directly to job behaviors.

Since training literature contains very little information about how to evaluate transfer of technical
training to job behaviors, a decision was made to survey training literature and to survey training professionals in other corporations. In several cases, benchmarks of third-level evaluation were the simplest possible, calling of supervisors after training to see if any changes had been observed. Several companies approached third-level evaluation in a more complex manner. One company considered its best practice to be taking a quality assurance approach to training, with trained observers in classes and a change committee reviewing work behaviors of trainees after training.

Other companies evaluated post-training behaviors in various ways. Some of the other practices reported were: (a) focus groups, (b) certifications, (c) expert judges who followed the sequence of training, practice, and final performance of trained skills, and (d) trainers who formed rotating teams to go out and inspect.

Although all companies who were interviewed or read about expressed interest and/or expertise in third-level evaluation, only one of the nineteen companies surveyed or reviewed yielded information on how to evaluate training transfer for technical training (i.e., training wherein knowledge and not skills were the focus). One organization doing third-level evaluation on technical training was found in the literature. The class being evaluated was a "knowledge class." Navy Personnel Research Department staff
persons reported stating course goals as competencies (Borman, Toquam, & Rosse, 1979). The same competencies were later evaluated by supervisors in conferences with trainees using a checksheet.

Since Physics/Chemistry classes contained only goals for learning and not for transfer of learning to the job, the building of a competency model seemed appropriate. Approximately fifty supervisors and technicians were telephoned to determine desirable behaviors which might be established or improved by technicians taking Physics/Chemistry classes. While serving as preliminary background data, the behaviors listed were later used to augment information gained in more formal procedures to build a competency model.

Early in the interviews, behaviors gathered from supervisors and technicians resembled the competence anchors on behaviorally anchored rating scales. Consequently, the researcher was influenced to develop behaviorally anchored rating scales to measure levels of competence before and after training.

Four groups of subjects were tested via the Solomon Four-Group Design (See Appendix A). The first two were experimental groups who received training in Physics/Chemistry. A group of prospective trainees comprised a first group whose supervisors responded to a pretest designed to measure certain behaviors of the
technicians on the job. Supervisors of the first group again responded to the same measurement instrument after technicians completed a Physics/Chemistry class. Both a posttest and a retrospective pretest were administered at that time. Subjects who had already received the training composed a second group. Their supervisors were asked to respond only to a posttest and a retrospective pretest designed to measure technicians' behaviors on the job.

Two non-treatment or control groups participated. One such group was comprised of technicians who were eligible, but had not actually taken a Physics/Chemistry class. Their supervisors evaluated with a pretest instrument at approximately the same time the experimental group received a pretest, and again in three months with a posttest and retrospective pretest. Another non-treatment or control group was evaluated only with a posttest and retrospective pretest.

**Development of Behaviorally Anchored Rating Scales**

As examples of competent behavior were gathered from supervisors and technicians, key areas or dimensions began to emerge. A group composed of two engineer/supervisors and a training manager with experience in semiconductor manufacturing met to list other dimensions. Six clearcut dimensions emerged and seemed to encompass all areas affected by training in Physics/Chemistry. They were: (a) Problem Solving, (b) Communication with Supervisor, (c)
Communication as Part of a Group, (d) Communication with Support Groups, (e) Troubleshooting, and (f) Understanding Total Process.

To collect critical incidents demonstrating competent use of focus dimensions, trainees in a Physics/Chemistry class were given time at various intervals throughout a three-day class to jot down ways in which they intended to use the training after they returned to their work. The projected behaviors were added to a growing list of competent behaviors (see Document 1, Appendix A).

To add to the list of critical incidents, the investigator met with a group of three engineers and technicians to try to exhaust the list of critical incidents and to place them under appropriate dimensions. The first sort revealed that another dimension was necessary, that of Calculations Involving Process.

In a second sort or re-translation with three engineers, many critical incidents were sorted into dimensions other than those for which they were originally intended. In cases of non-agreement, critical incidents were dropped. In addition, the last sorting prompted a final dimension to be added, that of Written Communication. In instances for which only excellent or average behaviors had been listed, the group extrapolated from the excellent behavior to versions of the behavior which were average,
less than average, and poor. All final placements were by unanimous agreement.

The final phase of development of the BARS consisted of six subject matter experts (in this case, engineer/supervisors) rating appropriateness of critical incidents for four succeeding levels of anchors under each dimension. Only four anchors were used due to the difficulty of eliciting many levels of each dimension. In some cases, even the four levels chosen indicated simply more or less of a particular behavior. Subject matter experts rated on a scale from one to four. Means and standard deviations were calculated. Anchors with high agreement and low variation were retained (see Appendix B).

For maximum discrimination power, the four levels of each dimension were spread over a ten-point scale. Least effective behavior was given rating potential of 1-2. Merely acceptable behavior was given a rating potential of 3-5. Effective behavior was scaled at 6-8. Highly effective behavior was given a rating potential of 9-10. Spreading the four levels widely over a scale of 10 was intended to give future raters the potential to rate as precisely as possible and to negate as much as possible the central tendency of raters. (They could, if compelled to rate centrally, choose the center of ranges at several levels.) As culmination of the BARS development, dimensions and rating scale were placed on a Scantron form to be used
by filling in with a pencil or pen. Although the instrument follows all procedures for development of behaviorally anchored rating scales, confusion among raters prompted a decision to use the phrase "behavior description scales."

**Administration of Evaluation Instrument**

In order to administer the BARS in a manner deemed appropriate, consideration was given to the total evaluation process. Existence of a needs assessment process was investigated. Although no formal needs assessment was in place at the time of project initiation, a formal management survey was later instituted. A promising vehicle was observed early on, however, in the form of design committees for technical training classes. The design committees commented freely about their desires for inclusion or exclusion of materials relating to specific training needs in their areas of responsibility.

Evaluation efforts also revealed that informal procedures are enacted each time a training manager visits informally with training managers at local sites, with managers at the various wafer fabs, and with engineers and technicians who call to discuss classes or to actually take classes. Consequently, the informal needs assessment is ongoing and dynamic, leading to development of courses to meet current directional emphases, finetuning of already-well-developed courses, and in certain cases, total changes of direction.
A decision was made to administer the BARS both to trainees and to their supervisors. Contacts were made directly with supervisors to explain the purpose of the project, to assure confidentiality, and to elicit their cooperation. The first groups were asked to respond prior to technicians taking Physics/Chemistry classes, and supervisors of technicians in the control group were asked to respond at the same time. After the telephone conversation, supervisors were mailed copies of the BARS and Scantron response forms. They were asked to mail them back in sealed envelopes via the internal mail system. Posttests and retrospective pretests were administered in the same manner to groups three months after training, or in the case of control groups, during the same time interval. Groups consisted of the four groups described earlier in the Solomon Four-Group Design.

In addition to supervisor evaluations, technician/trainees and controls were asked to self-evaluate. Contacting the technicians was very time-consuming, and in most cases, impossible without the assistance of supervisors. Consequently, technicians were contacted directly by telephone or personally when possible, and otherwise through their supervisors. In cases of supervisors being contacted and asked to deliver test materials to technicians, written instructions made clear the confidentiality and purpose of the study. Supervisors
were asked to assure technicians of confidentiality by explaining that the response forms could be mailed back in sealed packets.

Whether trainees or their supervisors evaluated, follow-up calls were a necessity. Phone dictation proved to be effective with some evaluators who were not pleased to be faced with what they considered paperwork. A follow-up luncheon was given to thank supervisors who participated in the evaluation and to gather late evaluations. The luncheons were also used to gather information and ideas which could contribute to an ongoing needs-assessment effort.
CHAPTER III

RESULTS

Means and standard deviations (see Table 1, Appendix E) were computed for each of the dependent variables (posttest dimensions 1-8). Correlation matrices (see Table 6, Appendix E) were run to determine correlations between the pretests, posttests, and retrospective pretests. Posttest--retrospective pretest correlations were found to be significant at the .05 level ($p < .0001$) in the case of all eight dimensions.

There was insufficient evidence (i.e., no statistically significant correlations) to reject the null hypothesis of no correlation between posttest and traditional pretest in the case of Dimensions 1 (Problem Solving), 4 (Written Communication), and 8 (Understanding Total Process). In the case of Dimensions 2 (Communication with Supervisor), 3 (Communication as Part of a Group), 5 (Communication with Support Groups), 6 (Troubleshooting), and 7 (Calculations involving Process), there was sufficient evidence to reject the null hypothesis of no correlation and to conclude that posttest and traditional pretest are correlated at the .05 level of significance ($r = .60746$, $p < .0001$; $r = .50373$, $p < .0006$; $r = .47652$, $p < .0012$; $r = .53098$, $p < .0002$; $r = .34778$, $p < .0223$, respectively).
Paired $t$ tests were computed on subjects' traditional pretest and retrospective pretest mean differences to determine if there were any significant differences between them (see Table 2, Appendix E). On Dimensions 1 (Problem Solving), 2 (Communication with Supervisor), 3 (Communication as Part of a Group), 4 (Written Communication), 5 (Communication with Support Groups), 7 (Calculations involving Process), and 8 (Understanding Total Process), there was insufficient evidence to reject the null hypothesis of mean difference being 0. Only for Dimension 6 (Troubleshooting) was there sufficient evidence to conclude that the mean difference between pretest and retroactive pretest was different than 0. The retrospective pretest score for Dimension 6 was significantly higher than that for the traditional pretest.

Due to unequal numbers of subjects in experimental groups, the general linear models procedure was used to compute an analysis of variance (ANOVA) on the dependent variables (posttest dimensions 1-8). Variances for treatment and pretest were analyzed to determine if any effects were significant (see Tables 3, Appendix E). No main effect was found for any variable. A significant interaction was found for only one variable. In the case of Posttest Dimension 5 (Communication with Support Groups), there was sufficient evidence ($p < .05$) to reject the null
hypothesis of no interaction effect and to conclude that a treatment-pretest interaction exists.

Figure 2. Pretest--Treatment Interaction for Dimension 5 (Communication with Support Group) Based on ANOVA

The general linear models procedure was used to compute an analysis of covariance (ANCOVA) on the dependent variables (posttest dimensions 1-8). The retrospective pretest scores, having been found highly correlated with posttests, were used as covariates. Worksites were utilized as covariates as well, with the significance of their linear relationship to posttest scores revealed in the ANCOVA analysis (see Table 4, Appendix E). Analysis of posttest scores revealed a treatment--pretest interaction effect,
significant at the .05 level of significance ($p < .05$) for Dimension 1 (Problem Solving). While the trained and pretested group scored lower than the trained group with no pretest, the pretested control group scored higher than the control group with no pretest.

**Figure 3.** Pretest--Treatment Interaction Effect for Dimension 1 (Problem Solving) Based on Analysis of Covariance

![Graph showing pretest and no pretest effects for trained and control groups.]

Main effects for pretest on Dimensions 1 (Problem Solving) and 2 (Communication with Supervisor) were significant at the .05 level of significance ($p < .05$). No main effects were observed for training.
The general linear models procedure was used to compute an analysis of covariance on posttest scores. Traditional pretests provided the covariate, and only the posttests for subjects having taken a traditional pretest were analyzed (see Table 5, Appendix E). No treatment effect was found to be significant.

Hypothesis 1

For technicians in semiconductor manufacturing companies, technicians receiving training in Physics/Chemistry classes demonstrate higher scores on the eight dimensions of the Behaviorally Description Scales for Semiconductor Technicians posttest, as rated by their supervisors, than technicians who do not receive the training. No statistically significant main effect for training was found for any of the dependent variables on the analysis of variance (ANOVA). Based on an analysis of covariance (ANCOVA), which utilized both retrospective pretest and worksites as covariates, again no effect for training was found. Lastly, no main effect for training was found in the analysis of covariance which tested significance of training on pretested groups only, using the traditional pretest as a covariate.

Hypothesis 2

Technicians whose supervisors rate them on a traditional pretest score the same on the posttests as technicians whose supervisors do not rate them on a
traditional pretest (i.e., no pretest effect). On the ANOVA (analysis of variance) procedure which utilized no covariates, a pretest-training interaction, significant at the .05 level of significance ($p < .05$) was found to exist for Dimension 5 (Communication with Support Groups). The trained group with no pretest scored higher on the posttest than did the trained and pretested group. The control group with no pretest scored lower on the posttest than did the pretested control group.

ANCOVA results revealed the same type of interaction, except for Dimension 1 (Problem Solving). Again, the trained group with no pretest scored higher on posttest than the trained and pretested group. The control group with no pretest scored lower on posttest than the pretested control group. Main effects for pretesting on Dimensions 1 and 2, revealed by the ANCOVA using retrospective pretest and worksites as covariates, were in favor of no pretesting.

Hypothesis 3

Supervisors rate technicians differently on a retrospective pretest than on a traditional pretest, thus demonstrating a type of response shift bias. Only on Dimension 6 (Troubleshooting) did the hypothesis hold true. In this one case, based on a paired $t$ test procedure, at a significance of .05 ($p < .05$), the retrospective pretest rating was significantly higher than the traditional pretest rating. On none of the other dimensions was there a
significant difference between pretest and retrospective pretest ratings.
CHAPTER IV

DISCUSSION

Based on the results found in the present study, no support exists for the hypothesis that technicians who receive training in Physics/Chemistry classes will score higher on posttest measures of the Behavior Description Scales (BDS) for Semiconductor Technicians than technicians who do not receive the training. Although small gains in the eight dimensions of the BDS were demonstrated for technicians who have taken the Physics/Chemistry classes, equivalent gains can be shown for technicians who have not taken the classes.

Several possible explanations exist for the apparent failure of technical training to manifest behavioral changes in the workplace. An inescapable conclusion resides in the possibility that technicians have not actually learned the concepts taught in the classes. While Level 2 evaluation mechanisms are in place in the company, utilization of learning evaluation are presently limited by time constraints of the training managers. A first step in promoting transfer of technical training to workplace behaviors must involve determination of concept mastery at the learning level. Modern technology makes possible
real-time evaluation of learning through audience response systems, an option being considered by the training department at the present time.

Several organizational factors come into play in evaluating workplace behaviors. Most of the behaviors identified by technicians and supervisors as being facilitated by knowledge of Physics/Chemistry are also emphasized, taught, and encouraged in additional manners. For example, one supervisor in the study warned that two control subjects whom she evaluated would show tremendous gains during the three-month evaluation period because they were new employees at the beginning of the study. In the interim, the new technicians received direct training and supervision from their supervisor. Direct one-on-one training is difficult to surpass.

Another organizational factor which almost certainly had a confounding effect was reported by an evaluating supervisor. He expressed an unsureness of what had caused his supervisee to improve in troubleshooting. The trainee had taken several classes which could have triggered increased competence. He was not certain which of the classes should have received credit for the accomplishment.

One organizational factor which impeded scientific rigor was discovered at the end of the study. A supervisor who served as a rater commented that he was the technician's supervisor, but not the direct supervisor. Unlike several
others in the same position, he had not turned the evaluation over to the supervisor most directly involved with the trainee.

Motivational factors seem to play a large role in many of the behaviors used to evaluate transfer of technical training. As Mager (1970) commented, an employee who cannot be forced to perform a task even at gunpoint is a prime candidate for training. Otherwise, lack of motivation or some organizational constraint is the cause of lack of performance.

A supervisor who evaluated both trainees and control subjects pointed out a situation which seems to demonstrate the importance of motivation. When asked why a control subject had improved as much as the trainees during the time period, the supervisor explained that two of the trainees were young, bright, and enthusiastic. When the motivated young technicians returned from Physics/Chemistry classes, they set out exuberantly to apply the principles learned. An older technician who did not like or want training (the control subject), then became highly motivated to show what he knew on the job. He did not want to be outperformed.

Reward factors may also play an important role in the extent to which classroom learning is transferred to workplace behaviors. One supervisor asked for copies of the projected uses of concepts which his technicians had listed during development of the Behavior Description Scales. He
wanted to be sure that he would recognize their efforts to apply the learnings. In cases where recognition is withheld, appropriate behaviors go unrewarded by the system.

The organizational elements which fail to facilitate transfer of technical training are ripe areas not so much for improving evaluation techniques, as for encouraging training managers to assume the role of organizational diagnosticians and consultants. Working with managers to improve organizational factors may greatly increase the effectiveness of training as an adjunct to other sturdy mechanisms.

A number of course presentation techniques can be devised to increase motivation to transfer learning. Again, the techniques are not as strictly within the parameters of course evaluation as in the realm of course development and organizational development. One example might be to send out to supervisors checksheets on which to check off behaviors on a weekly or monthly basis, once a supervisor approves enrollment of his or her supervised technicians in a course. All such devices would have increased chances of being effective if designed not to be time-consuming.

Course design would be improved by incorporating methods to promote the goals implied by the items on the behavior description scales. The items on the test instrument were elicited from supervisors and technicians as desired goals of taking physics/chemistry classes.
Consequently, the classes should teach not only scientific concepts, but how to use the concepts in the workplace. Behavioral transfer to the work setting would then be augmented by class design. Incorporation of methods directed toward behavioral transfer of scientific concepts would more surely meet the criteria of the content validity model (see Figure 1) espoused by Goldstein (1986), as well.

For both evaluation and course development purposes, an important enabler is support of management at the highest level possible. Knowledge that training is meeting the goals of management provides motivation and reward impetus to apply learned principles like nothing else can. In addition, management support paves the way for more effective and efficient means of evaluation, thus enabling the most vigorous application of scientific rigor needed to gather pertinent information.

Pretesting effects cannot be supported by the present study. A pretesting-training interaction effect was observed for Dimension 5 (Communication with Support Groups) in the ANOVA procedure. A pretesting-training interaction effect was observed for Dimension 1 (Problem Solving) in the ANCOVA procedure. Main effects were observed for Dimensions 1 and 2 (Communication with Supervisor) in the ANCOVA procedure. However, pretesting main effects on two out of eight dimensions and interaction effects on two out of eight
dimensions do not support the conclusion that pretesting had an overall effect.

Response shift was predicted for the evaluation of technical training. While a response shift bias on self-report measures is supported in the literature and the supervisors in this study did give self reports, their self reports were not about their own work. The response shift noted for Dimension 6 (Troubleshooting) was not in the supposed direction, as the mean retrospective pretest score was higher than the mean traditional pretest score. The indication may be that supervisors watched for and unintentionally promoted troubleshooting skills as a result of the increased awareness of being a rater. They may have become more aware of troubleshooting after learning that Physics/Chemistry classes are presumed by many to increase troubleshooting skills.

In light of the fact that no effect was found for training, conclusions about response shift bias are not convincing, no matter what the reasons may be. In addition, response shift bias was evident in the case of only one dimension. Logically, no response shift bias is expected in the absence of training effect. However, because no significant difference existed between pretests and retrospective pretests, a useful conclusion can be drawn. Apparently, training managers concerned about efficient use of resources would achieve as effective results by
administering retrospective pretests contiguous in time to the posttest as they would by administering the traditional pretests and posttests at two separate times. A large savings in time and cost could be realized by the more efficient means of testing.

Future Research

In future studies of training evaluation, the most prime and urgent recommendation that this study highlights is the need to first of all gather top management support. In the present study, top training management was eager for results. However, in some cases, much more cooperation could have been gained had top managers in the wafer fabrication units been brought into the study early rather than late. Access to workers would then be more readily attained. Access to the technicians themselves may well precipitate important findings in regard to retrospective pretesting. Whereas technicians who were within reach for the present study were enthusiastic and eager to share their perceptions, a limited number were available. It is the belief of the present researcher that technician availability would have been facilitated by top managers who wanted to learn if their goals were being met by training. Management surveys and interviews were instituted during the course of the evaluation projects, quite possibly due in part to the cry for a needs assessment from the present researcher and others. But, alas, management involvement in
the evaluation effort was not in time to augment the present evaluation undertaking.

Another recommendation for future research is to gather as many sources of information on intervening variables as possible. Analysis of data from the present study shows a slight improvement in teasing out training effects when covariates are added to analyses of variance. Just as Brinkerhoff (1989) and Dixon (1987) have stated, ferreting out the effects of training by itself may be a futile task in the area of workplace behavior.

While simple learning may be easily separable from other contributory factors, third-level or behavioral evaluation is much more complex. Many factors in the workplace may enable or encourage workers to use concepts learned in classes. All surveys and questionnaires, as well as the measurement instruments themselves should contain queries regarding demographics and other related variables. Such variables might include: (1) years of employment, (2) formal education level, (3) other trainings related to the same areas of investigation, (4) demonstrated competencies in specific areas, such as ratings or certifications, (5) performance evaluation data, and (6) workplace contingencies. For maximum utilization of effects of intervening variables, statistical expertise should be obtained early in the evaluation project to determine how
intervening variables may contribute to evaluation before the information is gathered.

In light of the lack of training effects and pretesting effects, as well as lack of response shift bias observed, evaluation of technical training at the behavioral level seems to be a difficult, if not impossible, task. Training managers who are eager to demonstrate effects of training would do well to begin the effort with evaluation of training for specific behaviors. Courses such as Physics/Chemistry probably belong in a series of classes to be evaluated with the understanding that they are foundations for behavior-oriented classes. The behavior-oriented classes should be the focus of evaluation.
APPENDIX A

PILOTING LEVEL III EVALUATION OF TECHNICAL TRAINING
VIA SOLOMON FOUR-GROUP DESIGN WITH
RETROSPECTIVE PRETEST
PILOTTING LEVEL III EVALUATION OF TECHNICAL TRAINING CLASSES VIA EXPERIMENTAL DESIGN

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>Class</th>
<th>Post/Retro Pre</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

E-1
Δ from Pre to Post presumably measures change due to classroom learning. But what else could have had an effect?
- Department Goals?
- Improved supervision or leadership?
- Self-improvement?
- Getting a new co-worker who is a super-high performer?

E-2
Did this group perform as well as E-1? What are the implications?
If not, what are the implications?

C-1
Did this group show as great a Δ as E-1? Or score as high on Post-test as E-2?
What are the implications?
If not, what are the implications?

C-2
Did this group perform as well as E-1?
Implications?
E-2? Implications?
C-2? Implications?
APPENDIX B

BEHAVIOR DESCRIPTION SCALES
Behavior Description Scales

**EVALUATION SCALE:** 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - 10

<table>
<thead>
<tr>
<th>EVALUATION RANGE</th>
<th>EVALUATION DESCRIPTION</th>
</tr>
</thead>
</table>

### I. PROBLEM SOLVING

1 - 2  | Often fails to recognize apparent problems, even when objective data are available. |
3 - 5  | Takes directions from someone else, tries 1 or 2 things, then comes back for additional directions. |
6 - 8  | Defines the specific problem and offers a solution. |
9 - 10 | Defines the specific problem and offers a solution. In addition, recognizes chronic problems and takes initiative to solve them. |

### II. COMMUNICATION WITH SUPERVISOR

1 - 2  | Rarely communicates technical information to supervisor. Even if asked, is unable to communicate technical information effectively. |
3 - 5  | Communicates technical information at a minimal level and only if asked. |
6 - 8  | If prompted, communicates appropriate technical information as required. |
9 - 10 | Communicates technical information as appropriate, without being prompted. |

### III. COMMUNICATION AS PART OF A GROUP

1 - 2  | Never offers input on the job or in technically-oriented meetings within his or her group. |
3 - 5  | While at work, makes occasional comments regarding the work of the group. |
6 - 8  | In technically-oriented meetings within his or her own group, offers input if asked. |
9 - 10 | Contributes to the team effort, often and appropriately offering input in technically-oriented meetings with his or her own group. |
Behavior Description Scales

<table>
<thead>
<tr>
<th>EVALUATION SCALE</th>
<th>EVALUATION RANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 2</td>
<td>3 - 5</td>
</tr>
</tbody>
</table>

IV. WRITTEN COMMUNICATION

1 - 2 Usually fails to convey needed technical information in written form.

3 - 5 Generates written reports which require a great deal of interpretation and questioning.

6 - 8 Produces written material which is somewhat helpful, but does not adequately describe the technical situation.

9 - 10 Communicates technical information effectively in written form, using appropriate details.

V. COMMUNICATION WITH SUPPORT GROUPS

1 - 2 Communicates very little with anyone outside his or her own organization.

3 - 5 Takes part in discussions with support group(s), but offers no expertise.

6 - 8 While sitting in meetings with various support groups, volunteers information relating to his or her own area of expertise.

9 - 10 Often communicates with other work groups to solve equipment / process problems.

VI. TROUBLESHOOTING

1 - 2 Unable to apply principles of chemistry and physics to carry out problem solving and troubleshooting.

3 - 5 Only rarely utilizes chemistry and physics concepts to troubleshoot process problems.

6 - 8 Systematically goes through steps to troubleshoot a problem, using the laws of chemistry and physics as they apply.

9 - 10 Utilizing all available resources, systematically and independently applies the laws of chemistry and physics to troubleshoot a problem.
Behavior Description Scales

EVALUATION SCALE  

\[1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10\]

VII. CALCULATIONS INVOLVING PROCESS

1 - 2  
Does not use calculations involving process for situations in which calculations are pertinent.

3 - 5  
Uses calculations involving process if explained thoroughly by his or her supervisor.

6 - 8  
Uses calculations involving process, asking precise questions about calculations and/or graphs when guidance is needed.

9 - 10  
Uses calculations involving process independently when appropriate.

VIII. UNDERSTANDING TOTAL PROCESS

1 - 2  
Demonstrates sufficient understanding of process to assist a technician with a different function, but only with very specific directions and constant supervision.

3 - 4  
Demonstrates sufficient understanding of process to assist a technician with a different function with close supervision.

6 - 8  
Demonstrates sufficient understanding of process to perform a function different than his or her own with very close supervision.

9 - 10  
Demonstrates sufficient understanding of process to perform a function different than his or her own with minimal supervision.
APPENDIX C

PROCESS FLOW CHART FOR DEVELOPMENT

BEHAVIOR DESCRIPTION SCALES
Process Flow Chart for Development
Behavior Description Scales

REQUEST FOR PROOF OF ADDED VALUE

1. INTERVIEW TRAINING PROFESSIONALS
   - BENCHMARK
   - IDENTIFIED PROTOTYPE
   - SELECT APPROPRIATE MODEL

2. REVIEW LITERATURE

3. INTERVIEW TRAINEES AND SUPERVISORS
   - COMPETENCY MODEL

IV. GROUP OF 3-6 SME'S SELECT DIMENSIONS

V. ELICIT PROPOSED USES OF TRAINING FROM TRAINEES

VI. LIST OF VARIOUS CRITICAL INCIDENTS
   - GROUP OF 3-6 SME'S LIST CRITICAL INCIDENTS
   - LIST OF CRITICAL INCIDENTS

COMBINE CRITICAL INCIDENTS AND "TRAINEE" BEHAVIORS

LIST OF BEHAVIORS
GROUP OF 3-6 SME'S SORT INCIDENTS UNDER DIMENSIONS

DISCARD INCIDENTS WITH LOW AGREEMENT

DIMENSIONS WITH 5-10 CRITICAL INCIDENTS EACH

6 - 10 SUPERVISORS RANK CRITICAL INCIDENTS

CALCULATE x̄ and S∀S

SELECT INCIDENTS WHICH HAVE LEAST VARIANCE FOR LEVELS

DIMENSIONS WITH BEHAVIOR FOR LEVELS 1, 2, 3, 4, etc.

SPREAD BEHAVIOR OVER 10 POINT SCALE

PLACE ITEMS ON SCANTRON

* BARS ON SCANTRON

ADMINISTER VIA EXPERIMENTAL DESIGN TO TREATMENT (TRAINED) AND CONTROL GROUP

ANALYZE RESULTS

INSTRUMENT VALIDATED VIA EXPERIMENTAL DESIGN AND SUPERVISION CRITERION

* Behaviorally anchored rating scale
APPENDIX D

PROCESS FLOW CHART FOR ADMINISTRATION OF BEHAVIOR DESCRIPTION SCALES
Process Flow Chart for Administration of Behavior Description Scales

1. NEEDS ASSESSMENT

2. CUSTOMER CRITIQUE OF DESIGN

3. DEVELOP EVALUATION INSTRUMENT

4. DISCUSS PURPOSE WITH TECHNICIANS

5. SUPERVISOR EVALUATE PERFORMANCE

6. CALL SUPERVISOR TO EXPLAIN EVALUATION

- SEND BARS TO SUPERVISORS
- SUPERVISORS EVALUATE TECHNICIANS
- PRE-TEST EVALUATION

7. TECHNICIANS LIST PROPOSED USES DURING CLASS

8. LIST OF PROPOSED USES

9. 1 COPY TO TECH
   1 COPY TO SUPERVISOR

10. LIST OF PROPOSED BEHAVIORS FOR S'S & T'S
WAIT 2 MONTHS TO GIVE TIME FOR BEHAVIORAL CHANGES

IX.
SUPERVISOR EVALUATE PERFORMANCE

X.
POST AND RETRO EVALUATION WITH SUPERVISORS

SEND PACKETS

FOLLOW UP WITH CALLS OR MSGS. ALLOW DICTATION IF NECESSARY

POST-TEST AND RETROSPECTIVE PRE-TEST EVALUATION

TABULATE RESULTS

XII.
DISTRIBUTE PACKETS TO TECHNICIANS THROUGH SUPERVISORS

FOLLOW UP WITH CALLS. ALLOW PHONE DICTATION IF NECESSARY

POST-TEST AND RETROSPECTIVE PRE-TEST EVALUATION

XIII.
CONDUCT INTERVIEWS TO CLARIFY CAUSES

PERCENT TRANSFER AS INDICATOR FOR DIRECTION OF FUTURE DEVELOPMENT AND/OR DELIVERY

KNOWLEDGE OF NEEDED CHANGES

XIV.

PERCENT CHANGE MEETS CRITERIA?

NO

YES

XV.
REPORT RESULTS
APPENDIX E

STATISTICAL TABLES
Table 1

Posttest Means and Standard Deviations for Dimensions 1-8

Dimension 1: Problem Solving

<table>
<thead>
<tr>
<th>GROUP</th>
<th>PRETEST</th>
<th>PST1 MEAN</th>
<th>STD</th>
<th>N</th>
<th>PRE1 MEAN</th>
<th>STD</th>
<th>N</th>
<th>RET1 MEAN</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Pretest Taken</td>
<td>22.00</td>
<td>6.77</td>
<td>2.29</td>
<td>21.00</td>
<td>6.05</td>
<td>2.96</td>
<td>22.00</td>
<td>6.45</td>
</tr>
<tr>
<td></td>
<td>Pretest not Taken</td>
<td>27.00</td>
<td>6.22</td>
<td>2.31</td>
<td>6.00</td>
<td></td>
<td>1.96</td>
<td>31.00</td>
<td>7.13</td>
</tr>
<tr>
<td>Experimental</td>
<td>Pretest Taken</td>
<td>23.00</td>
<td>6.35</td>
<td>1.21</td>
<td>22.00</td>
<td>6.64</td>
<td>2.32</td>
<td>23.00</td>
<td>7.09</td>
</tr>
<tr>
<td></td>
<td>Pretest not Taken</td>
<td>11.00</td>
<td>7.19</td>
<td>1.69</td>
<td>6.00</td>
<td></td>
<td>1.96</td>
<td>31.00</td>
<td>6.77</td>
</tr>
</tbody>
</table>

Dimension 2: Communication with Supervisor

<table>
<thead>
<tr>
<th>GROUP</th>
<th>PRETEST</th>
<th>PST2 MEAN</th>
<th>STD</th>
<th>PRE2 MEAN</th>
<th>STD</th>
<th>RET2 MEAN</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Pretest Taken</td>
<td>22.00</td>
<td>7.69</td>
<td>2.60</td>
<td>21.00</td>
<td>6.10</td>
<td>3.13</td>
</tr>
<tr>
<td></td>
<td>Pretest not Taken</td>
<td>27.00</td>
<td>6.69</td>
<td>1.78</td>
<td>0.00</td>
<td></td>
<td>1.96</td>
</tr>
<tr>
<td>Experimental</td>
<td>Pretest Taken</td>
<td>23.00</td>
<td>7.65</td>
<td>1.99</td>
<td>22.00</td>
<td>7.41</td>
<td>2.88</td>
</tr>
<tr>
<td></td>
<td>Pretest not Taken</td>
<td>31.00</td>
<td>7.68</td>
<td>1.96</td>
<td>6.00</td>
<td></td>
<td>1.96</td>
</tr>
</tbody>
</table>

(table continues)
Dimension 3: Communication as Part of a Group

<table>
<thead>
<tr>
<th>GROUP</th>
<th>PST3</th>
<th>PRE3</th>
<th>RET3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>MEAN</td>
<td>STD</td>
</tr>
<tr>
<td>PRETEST</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Taken</td>
<td>22.00</td>
<td>7.14</td>
<td>2.08</td>
</tr>
<tr>
<td>Pretest not</td>
<td>27.00</td>
<td>6.37</td>
<td>1.84</td>
</tr>
<tr>
<td>Experimental</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Taken</td>
<td>23.00</td>
<td>7.04</td>
<td>2.51</td>
</tr>
<tr>
<td>Pretest not</td>
<td>31.00</td>
<td>6.48</td>
<td>2.84</td>
</tr>
</tbody>
</table>

Dimension 4: Written Communication

<table>
<thead>
<tr>
<th>GROUP</th>
<th>PST4</th>
<th>PRE4</th>
<th>RET4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>MEAN</td>
<td>STD</td>
</tr>
<tr>
<td>PRETEST</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Taken</td>
<td>22.00</td>
<td>6.82</td>
<td>2.15</td>
</tr>
<tr>
<td>Pretest not</td>
<td>27.00</td>
<td>6.26</td>
<td>2.43</td>
</tr>
<tr>
<td>Experimental</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Taken</td>
<td>23.00</td>
<td>6.39</td>
<td>2.61</td>
</tr>
<tr>
<td>Pretest not</td>
<td>31.00</td>
<td>6.68</td>
<td>2.14</td>
</tr>
</tbody>
</table>

Dimension 5: Communication with Support Groups

<table>
<thead>
<tr>
<th>GROUP</th>
<th>PST5</th>
<th>PRE5</th>
<th>RET5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>MEAN</td>
<td>STD</td>
</tr>
<tr>
<td>PRETEST</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Taken</td>
<td>22.00</td>
<td>7.41</td>
<td>1.53</td>
</tr>
<tr>
<td>Pretest not</td>
<td>27.00</td>
<td>6.52</td>
<td>1.58</td>
</tr>
<tr>
<td>Experimental</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest Taken</td>
<td>23.00</td>
<td>7.04</td>
<td>2.01</td>
</tr>
<tr>
<td>Pretest not</td>
<td>31.00</td>
<td>7.65</td>
<td>1.89</td>
</tr>
</tbody>
</table>

(table continues)
Dimension 6: Troubleshooting

<table>
<thead>
<tr>
<th>GROUP</th>
<th>PST6</th>
<th>PRE6</th>
<th>RET6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>MEAN</td>
<td>STD</td>
</tr>
<tr>
<td>Control</td>
<td>Pretest Taken</td>
<td>22.00</td>
<td>6.65</td>
</tr>
<tr>
<td></td>
<td>Pretest not Taken</td>
<td>27.00</td>
<td>6.33</td>
</tr>
<tr>
<td>Experimental</td>
<td>Pretest Taken</td>
<td>23.00</td>
<td>6.43</td>
</tr>
<tr>
<td></td>
<td>Pretest not Taken</td>
<td>31.00</td>
<td>6.48</td>
</tr>
</tbody>
</table>

Dimension 7: Calculations involving Process

<table>
<thead>
<tr>
<th>GROUP</th>
<th>PST7</th>
<th>PRE7</th>
<th>RET7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>MEAN</td>
<td>STD</td>
</tr>
<tr>
<td>Control</td>
<td>Pretest Taken</td>
<td>22.00</td>
<td>6.35</td>
</tr>
<tr>
<td></td>
<td>Pretest not Taken</td>
<td>27.00</td>
<td>6.07</td>
</tr>
<tr>
<td>Experimental</td>
<td>Pretest Taken</td>
<td>23.00</td>
<td>5.78</td>
</tr>
<tr>
<td></td>
<td>Pretest not Taken</td>
<td>31.00</td>
<td>6.42</td>
</tr>
</tbody>
</table>

Dimension 8: Understanding Total Process

<table>
<thead>
<tr>
<th>GROUP</th>
<th>PST8</th>
<th>PRE8</th>
<th>RET8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>MEAN</td>
<td>STD</td>
</tr>
<tr>
<td>Control</td>
<td>Pretest Taken</td>
<td>22.00</td>
<td>6.64</td>
</tr>
<tr>
<td></td>
<td>Pretest not Taken</td>
<td>27.00</td>
<td>7.00</td>
</tr>
<tr>
<td>Experimental</td>
<td>Pretest Taken</td>
<td>23.00</td>
<td>6.39</td>
</tr>
<tr>
<td></td>
<td>Pretest not Taken</td>
<td>31.00</td>
<td>7.48</td>
</tr>
</tbody>
</table>

Note. PST = Posttest; PRE = Pretest; RET = Retrospective Pretest.
Table 2

**Paired T Tests on Pretests and Retrospective Pretests**

| Variable | N  | Mean   | Std Error | T     | Prob>|I| |
|----------|----|--------|-----------|-------|------|-----|
| DIF1     | 43 | -0.4186047 | 0.4159507 | -1.0063804 | 0.3200 |
| DIF2     | 43 | 0.2093823  | 0.3113389 | 0.6721554  | 0.5052 |
| DIF3     | 43 | -0.2790698 | 0.4406348 | -0.6333358 | 0.5299 |
| DIF4     | 43 | -0.2093823 | 0.4794309 | -0.4365641 | 0.6647 |
| DIF5     | 43 | -0.3488372 | 0.3258581 | -1.0705187 | 0.2905 |
| DIF6     | 43 | -0.6511628 | 0.3137381 | -2.0754981 | 0.0441 |
| DIF7     | 43 | -0.0232558 | 0.3742669 | 0.0621370  | 0.9507 |
| DIF8     | 43 | -0.3255814 | 0.4595738 | -0.7084420 | 0.4826 |
Table 3

Analysis of Variance Results with Significance

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>F Value</th>
<th>Probability &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dimension 1: Troubleshooting</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>0.62</td>
<td>0.3108</td>
</tr>
<tr>
<td>Pretest</td>
<td>0.27</td>
<td>0.6043</td>
</tr>
<tr>
<td>Group-Pretest</td>
<td>2.86</td>
<td>0.0940</td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dimension 2: Communication with Supervisor</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>2.67</td>
<td>0.1052</td>
</tr>
<tr>
<td>Pretest</td>
<td>0.05</td>
<td>0.8309</td>
</tr>
<tr>
<td>Group-Pretest</td>
<td>0.08</td>
<td>0.7837</td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dimension 3: Communication as Part of a Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>0.00</td>
<td>0.9827</td>
</tr>
<tr>
<td>Pretest</td>
<td>1.96</td>
<td>0.1641</td>
</tr>
<tr>
<td>Group-Pretest</td>
<td>0.05</td>
<td>0.8277</td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dimension 4: Written Communication</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>0.00</td>
<td>0.9925</td>
</tr>
<tr>
<td>Pretest</td>
<td>0.09</td>
<td>0.7690</td>
</tr>
<tr>
<td>Group-Pretest</td>
<td>0.83</td>
<td>0.3638</td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

_table continues_
<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
<th>Group</th>
<th>Pretest</th>
<th>Group-Pretest Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Communication with Support Groups</td>
<td>1.17</td>
<td>0.17</td>
<td>4.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Troubleshooting</td>
<td>0.01</td>
<td>0.14</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Calculations involving Process</td>
<td>0.22</td>
<td>0.03</td>
<td>1.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Understanding Total Process</td>
<td>0.07</td>
<td>2.66</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Group Pretest Group-Pretest Interaction
Table 4

Analysis of Covariance Results with Significance (using Retrospective Pretest and Worksites as Covariates)

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>F Value</th>
<th>Probability &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dimension 1: Problem Solving</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>0.02</td>
<td>0.8801</td>
</tr>
<tr>
<td>Pretest</td>
<td>3.95</td>
<td>0.0500</td>
</tr>
<tr>
<td>Group-Pretest</td>
<td>6.16</td>
<td>0.0149</td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dimension 2: Communication with Supervisor</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>0.58</td>
<td>0.5591</td>
</tr>
<tr>
<td>Pretest</td>
<td>4.71</td>
<td>0.0326</td>
</tr>
<tr>
<td>Group-Pretest</td>
<td>0.00</td>
<td>0.9490</td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dimension 3: Communication as Part of a Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>0.02</td>
<td>0.9962</td>
</tr>
<tr>
<td>Pretest</td>
<td>0.14</td>
<td>0.7119</td>
</tr>
<tr>
<td>Group-Pretest</td>
<td>0.67</td>
<td>0.4141</td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dimension 4: Written Communication</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>0.28</td>
<td>0.5990</td>
</tr>
<tr>
<td>Pretest</td>
<td>0.27</td>
<td>0.6030</td>
</tr>
<tr>
<td>Group-Pretest</td>
<td>1.66</td>
<td>0.2015</td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*table continues*
### Dimension 5: Communication with Support Groups

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Pretest</th>
<th>Group-Pretest Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.16</td>
<td>0.22</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>0.2848</td>
<td>0.6371</td>
<td>0.3139</td>
</tr>
</tbody>
</table>

### Dimension 6: Troubleshooting

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Pretest</th>
<th>Group-Pretest Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.06</td>
<td>2.17</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>0.1549</td>
<td>0.1438</td>
<td>0.5168</td>
</tr>
</tbody>
</table>

### Dimension 7: Calculations involving Process

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Pretest</th>
<th>Group-Pretest Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.00</td>
<td>0.53</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>0.9917</td>
<td>0.4705</td>
<td>0.5851</td>
</tr>
</tbody>
</table>

### Dimension 8: Understanding Total Process

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Pretest</th>
<th>Group-Pretest Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.8965</td>
<td>0.9405</td>
<td>0.9129</td>
</tr>
</tbody>
</table>
Table 5

Analysis of Covariance on Pretested Groups Using Pretest as Covariate

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Variance Source</th>
<th>F Value</th>
<th>Probability &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Group</td>
<td>0.30</td>
<td>0.5881</td>
</tr>
<tr>
<td>2</td>
<td>Group</td>
<td>0.08</td>
<td>0.7800</td>
</tr>
<tr>
<td>3</td>
<td>Group</td>
<td>0.34</td>
<td>0.5624</td>
</tr>
<tr>
<td>4</td>
<td>Group</td>
<td>0.13</td>
<td>0.7218</td>
</tr>
<tr>
<td>5</td>
<td>Group</td>
<td>2.01</td>
<td>0.1635</td>
</tr>
<tr>
<td>6</td>
<td>Group</td>
<td>0.48</td>
<td>0.4927</td>
</tr>
<tr>
<td>7</td>
<td>Group</td>
<td>0.95</td>
<td>0.3343</td>
</tr>
</tbody>
</table>
Table 6

Correlation Matrices for Pretest, Posttest, and Retrospective Pretest Scores

Dimension 1: Problem Solving

<table>
<thead>
<tr>
<th></th>
<th>PRE1</th>
<th>PST1</th>
<th>RET1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRE1</td>
<td>1.00000</td>
<td>0.11729</td>
<td>0.40783</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.4538</td>
<td>0.0066</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>43</td>
<td>100</td>
</tr>
<tr>
<td>PST1</td>
<td>0.11729</td>
<td>1.00000</td>
<td>0.61583</td>
</tr>
<tr>
<td></td>
<td>0.4538</td>
<td>0.0</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>103</td>
<td>100</td>
</tr>
<tr>
<td>RET1</td>
<td>0.40783</td>
<td>0.61583</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.0066</td>
<td>0.0001</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Dimension 2: Communication with Supervisor

<table>
<thead>
<tr>
<th></th>
<th>PRE2</th>
<th>PST2</th>
<th>RET2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRE2</td>
<td>1.00000</td>
<td>0.60746</td>
<td>0.69596</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.746</td>
<td>0.696</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>43</td>
<td>43</td>
</tr>
<tr>
<td>PST2</td>
<td>0.60746</td>
<td>1.00000</td>
<td>0.84883</td>
</tr>
<tr>
<td></td>
<td>0.746</td>
<td>0.0</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>103</td>
<td>100</td>
</tr>
<tr>
<td>RET2</td>
<td>0.69596</td>
<td>0.84883</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.696</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Dimension 3: Communication as Part of a Group

<table>
<thead>
<tr>
<th></th>
<th>PRE3</th>
<th>PST3</th>
<th>RET3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRE3</td>
<td>1.00000</td>
<td>0.50373</td>
<td>0.35604</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.504</td>
<td>0.356</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>43</td>
<td>43</td>
</tr>
<tr>
<td>PST3</td>
<td>0.50373</td>
<td>1.00000</td>
<td>0.76963</td>
</tr>
<tr>
<td></td>
<td>0.504</td>
<td>0.0</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>103</td>
<td>100</td>
</tr>
<tr>
<td>RET3</td>
<td>0.35604</td>
<td>0.76963</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.356</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Dimension 4: Written Communication

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / Number of Observations

<table>
<thead>
<tr>
<th></th>
<th>PRE4</th>
<th>PST4</th>
<th>RET4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRE4</td>
<td>1.00000</td>
<td>0.20307</td>
<td>0.29145</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.1915</td>
<td>0.0579</td>
</tr>
<tr>
<td>PST4</td>
<td>0.20307</td>
<td>1.00000</td>
<td>0.83204</td>
</tr>
<tr>
<td></td>
<td>0.1915</td>
<td>0.8</td>
<td>0.0001</td>
</tr>
<tr>
<td>RET4</td>
<td>0.29145</td>
<td>0.83204</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.0579</td>
<td>0.0001</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>43</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>103</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>103</td>
<td>100</td>
</tr>
</tbody>
</table>

Dimension 5: Communication with Support Groups

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / Number of Observations

<table>
<thead>
<tr>
<th></th>
<th>PRE5</th>
<th>PST5</th>
<th>RET5</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRE5</td>
<td>1.00000</td>
<td>0.47652</td>
<td>0.51642</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.0012</td>
<td>0.0004</td>
</tr>
<tr>
<td>PST5</td>
<td>0.47652</td>
<td>1.00000</td>
<td>0.87921</td>
</tr>
<tr>
<td></td>
<td>0.0012</td>
<td>0.0</td>
<td>0.0001</td>
</tr>
<tr>
<td>RET5</td>
<td>0.51642</td>
<td>0.87921</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.0004</td>
<td>0.0001</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>103</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Dimension 6: Troubleshooting

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / Number of Observations

<table>
<thead>
<tr>
<th></th>
<th>PRE6</th>
<th>PST6</th>
<th>RET6</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRE6</td>
<td>1.00000</td>
<td>0.53098</td>
<td>0.53496</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.0002</td>
<td>0.0002</td>
</tr>
<tr>
<td>PST6</td>
<td>0.53098</td>
<td>1.00000</td>
<td>0.73646</td>
</tr>
<tr>
<td></td>
<td>0.0002</td>
<td>0.0</td>
<td>0.0001</td>
</tr>
<tr>
<td>RET6</td>
<td>0.53496</td>
<td>0.73646</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.0002</td>
<td>0.0001</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>103</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

(table continues)
### Dimension 7: Calculations involving Process

<table>
<thead>
<tr>
<th></th>
<th>PRE7</th>
<th>PST7</th>
<th>RET7</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRE7</td>
<td>1.00000</td>
<td>0.34778</td>
<td>0.54272</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.0223</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>43</td>
<td>43</td>
</tr>
<tr>
<td>PST7</td>
<td>0.34778</td>
<td>1.00000</td>
<td>0.83739</td>
</tr>
<tr>
<td></td>
<td>0.0223</td>
<td>0.0</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>103</td>
<td>100</td>
</tr>
<tr>
<td>RET7</td>
<td>0.54272</td>
<td>0.83739</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.0002</td>
<td>0.0001</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

### Dimension 8: Understanding Total Process

<table>
<thead>
<tr>
<th></th>
<th>PRE8</th>
<th>PST8</th>
<th>RET8</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRE8</td>
<td>1.00000</td>
<td>0.10472</td>
<td>0.30005</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.5039</td>
<td>0.0506</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>43</td>
<td>43</td>
</tr>
<tr>
<td>PST8</td>
<td>0.10472</td>
<td>1.00000</td>
<td>0.75127</td>
</tr>
<tr>
<td></td>
<td>0.5039</td>
<td>0.0</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>103</td>
<td>100</td>
</tr>
<tr>
<td>RET8</td>
<td>0.30005</td>
<td>0.75127</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.0506</td>
<td>0.0001</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**Note.** PST = Posttest; PRE = Pretest; RET = Retrospective Pretest.
REFERENCES


of personnel programs.' *Journal of Applied Psychology*, 75, 410-417.


