APPLYING THE INTERNAL REFERENCING STRATEGY TO THE EVALUATION OF TRANSFER OF TRAINING IN FIELD SETTINGS

by

Daniel Jacob Watola

B.S., United States Air Force Academy, 1993

A thesis submitted to the

University of Colorado at Denver

in partial fulfillment

of the requirements for the degree of

Master of Arts

Psychology

1997

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	18 JUN 97	

4. TITLE AND SUBTITLE
APPLYING THE INTERNAL REFERENCING STRATEGY TO THE EVALUATION OF TRANSFER OF TRAINING IN FIELD SETTINGS

5. FUNDING NUMBERS

6. AUTHOR(S)

DANIEL JACOB WATOLA

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

UNIVERSITY OF COLORADO AT DENVER

8. PERFORMING ORGANIZATION REPORT NUMBER

97-068

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

DEPARTMENT OF THE AIR FORCE
AFIT/CI BLDG 125
2950 P STREET
WRIGHT-PATTERSON AFB OH 45433-7765

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

14. SUBJECT TERMS

15. NUMBER OF PAGES
94

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

This thesis for the Master of Arts

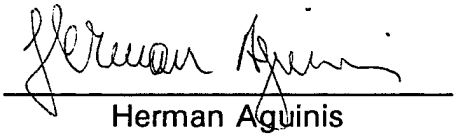
degree by

Daniel Jacob Watola

has been approved

by

Kurt Kraiger



Herman Aguinis

Brent G. Wilson

5/7/97

Date

Watola, Daniel Jacob (M.A., Psychology)

Applying the Internal Referencing Strategy to the Evaluation of Transfer of Training in Field Settings

Thesis directed by Professor Kurt Kraiger

ABSTRACT

Training evaluation is the determination of the effectiveness of a training program. While training researchers and practitioners generally prefer the use of an experimental research design to infer training effectiveness, practical constraints inherent in field settings often require that they employ a less rigorous quasi-experimental design. This paper evaluates a managerial training program in a field setting using the Internal Referencing Strategy (IRS), a quasi-experimental research design that infers training effectiveness when trainee pretest-posttest change on training-relevant test items is greater than pretest-posttest change on training-irrelevant test items. One hundred and eighty-two managers experienced a managerial training course and provided pretest and posttest data. Trainee learning was assessed using a 20-item multiple

choice knowledge test and trainee transfer of training was assessed using a 15-item behavioral questionnaire. An analysis of the evaluative data provided some evidence of learning, and demonstrated transfer of training after items exhibiting ceiling effects were excluded from the analysis. Implications regarding the use of the IRS approach in training evaluation are discussed.

This abstract accurately represents the content of the candidate's thesis. I recommend its publication.

Signed 
Kurt Kraiger

DEDICATION

This thesis is dedicated to my family and friends--thanks for supplying the ingredient that makes a good thesis truly great.

ACKNOWLEDGMENTS

A standing ovation to the members of my thesis committee: Kurt Kraiger, Herman Aguinis, and Brent Wilson. Kurt knows the rich and complex world of field research, and now I do too. Herman made difficult statistics easy, and easy statistics even easier. Brent kept me honest with a fresh perspective.

A hand salute to the U.S. Air Force for funding my graduate degree program, and a wink and a nod to several anonymous employees of a major regional telecommunications company.

CONTENTS

CHAPTER 1

INTRODUCTION

There has been an unparalleled growth in the number of training and development programs witnessed in organizational settings over the past several decades (Bunker & Cohen, 1977; Tannenbaum & Woods, 1992). This growth prompted Mathieu, Tannenbaum, and Salas (1992) to conclude that "nearly all employees will receive some form of training during their careers" (p. 828). Ralphs and Stephan's (1986) survey of human resource professionals in Fortune 500 companies appeared to support this conclusion, as respondents revealed that 91% of firms provided middle manager training, 75% provided sales training, 56% provided secretarial or clerical training, 51% provided executive development training, and 44% provided technical training.

Ford, Quinones, Sego, and Sorra (1992) attributed some of the growth in organizational training to increasing competitive pressures, as one mechanism through which organizations can improve the quality of their products and services is by improving the knowledge levels and technical skills of their employees. Technological growth, governmental

requirements, and the rise of the job enrichment movement have also been cited as contributors to the "industrial training boom" (Bunker & Cohen, 1977). Regardless of the reasons, this increase in training has been costly. The American Society for Training and Development estimated that organizations spend $210 billion each year on formal and informal employee training (Finkel, 1987); Wexley and Latham (1991) estimated training costs at approximately $2 trillion per decade.

In keeping with past and present trends in industrial training, there is some reason to believe that organizations will continue to rely on training and development programs in the future. Specifically, Goldstein and Gilliam (1990) identified four trends that are characteristic of the evolving, future workplace and are likely to precipitate continued organizational training: (a) changing workforce demographics, (b) increasing technological sophistication, (c) the shift from a manufacturing-orientation to a service-orientation, and (d) the emergence of a "global economy."

First, workforce demographics are in flux; the proportion of entry-level youth in the workforce is declining, and the proportion of older workers, women, racial minorities, and the undereducated is increasing (Cascio & Zammuto, 1989; Fullerton, 1989, 1995). Older workers may

require additional training as a result of job or career shifts, and to keep pace with technological change (Goldstein, 1980). The diversification of the workforce is likely to promote continued support for diversity training within organizations (Thayer, 1997). In addition to providing skills training to entry-level youth, organizations may also have to provide basic education in reading, writing, and mathematics to undereducated employees (Goldstein, 1993). Gorman (1988) estimated that half of the Fortune 500 firms, as "educators of last resort," spent more than $300 million each year on employee remedial education. For example, in just one year, Polaroid Corporation spent $700,000 on courses in English and math for 1000 new and veteran employees, American Express taught its new workers basic English and social skills at a cost of $10 million, and General Motors spent more than $25.5 million on employee remedial education programs.

Second, increasing technological sophistication is placing greater demands on those workers who must design, operate, and maintain advanced systems. As a result, training requirements are likely to grow in number, increase in complexity, and require frequent revisions, as organizations field still more advanced systems that effectively render previous training obsolete (Goldstein & Gilliam, 1990; Turnage, 1990).

Third, the economy's change from a manufacturing-orientation to a service-orientation requires that workers shift from working primarily with objects to working primarily with people (Klein & Hall, 1988). Thus, displaced manufacturers may need to learn interpersonal skills in order to succeed in the service industries (Goldstein & Gilliam, 1990).

Finally, organizational training is growing in response to the emergence of a global economy. Company employees that work closely with foreign nationals may need to learn the language, customs, and business practices of their foreign counterparts if they are to succeed (Thayer, 1997). Global competition continues to pressure organizations to improve the quality of their products and services. In light of the decreasing size of the overall workforce, organizations will have to find ways to do more with less, by maximizing the potential of the individual worker (Goldstein & Gilliam, 1990). This will require that organizations provide training to groups that have previously suffered from unfair discrimination, such as unskilled youth, older workers, women, and racial minorities (Cascio, 1991).

In view of the changes organizations will face in the coming years, there can be little doubt that the demand for training programs will continue to rise, as "training and development are used extensively as

strategies for confronting, coping, and adapting to change, and as strategies for improving job performance and organizational effectiveness" (Cascio, 1991, p. 360).

Given their past, present, and future, training and development programs can be described as "pervasive, expensive, and strategically important" (Tannenbaum & Woods, 1992, p. 63). If organizations are spending billions of dollars each year to change their employees' knowledge, behavior, and attitudes, then it is only logical that they should evaluate the effectiveness of their training programs. Organizations can rely on training evaluation to determine the degree to which a training program has achieved its intended objectives (Geber, 1995; Kirkpatrick, 1976; Newstrom, 1978); to provide feedback to trainees, trainers, and program designers in order to continuously improve the program (Blaiwes, Puig, & Regan, 1973; Newstrom, 1978; Swierczek & Carmichael, 1985); to make decisions about selecting, retaining, modifying, or eliminating programs (Geber, 1995; Grove & Ostroff, 1990; Kirkpatrick, 1978; Sackett & Mullen, 1993); and for legal or marketing documentation purposes (Grove & Ostroff, 1990; Sackett & Mullen, 1993).

Yet, there is evidence that organizations do not conduct extensive evaluations of their training programs. Ralphs and Stephan's (1986)

survey of the training practices of Fortune 500 companies revealed that 86% of organizations' training evaluations consisted of little more than trainee reactions obtained at the end of the course. Few efforts were made to assess changes in trainees' performance on the job, where the organization would learn if the training program's objectives had been met, or if modifications to the training program were necessary. Similarly, Saari, Johnson, McLaughlin, and Zimmerle (1988) examined the management training practices of 1,000 randomly selected U.S. companies with at least 1,000 employees and concluded the following:

> In general, there is little evidence of systematic evaluations of management training by U.S. companies. To the extent that evaluations are conducted, the primary method used is evaluation forms administered after program participation. . . . evaluations of management training, other than reactions of participants following program attendance, are not evident in U.S. companies. (p. 741-742)

Researchers have only recently grown more optimistic with respect to the future of training evaluation in organizations. Goldstein (1993) claimed that organizations are becoming increasingly interested in training evaluation as a means of determining if their training goals are being attained. Geber (1995) echoed this belief and attributed this growing interest to at least three factors. First, in a time of shrinking budgets, managers are beginning to demand that training departments provide

concrete evidence that their programs are both changing employees' behavior on the job and contributing to the bottom line. Second, the quality movement's emphasis on measurement has prompted training departments to discard such training evaluation measures as "number of satisfied trainees" and "total training hours delivered" in exchange for more meaningful measures that promote effective decision-making. Finally, technological advances have eased the burden of data collection and analysis, permitting training departments to apply more sophisticated evaluation strategies.

In sum, while evaluation should be an essential part of organizational training, only rarely do organizations engage in any rigorous form of training evaluation. In a time of tight training budgets, organizations that fail to evaluate their training programs will be unable to determine if their training dollars are being spent effectively (Ostroff, 1991). Lacking evaluative data, training decisions are likely to be based on anecdotes, reactions, hunches, or inertia (Tannenbaum & Woods, 1992). Therefore, training researchers and practitioners should seek to identify or develop rigorous, yet easily implemented, methods for generating useful evaluative data.

This thesis examines Haccoun and Hamtiaux's (1994) Internal Referencing Strategy (IRS) as a means of facilitating training evaluation in organizational settings. In order to understand the IRS approach and this study's research questions, it is important to understand several issues including training, both in terms of learning in the training context and the transfer of training to the job context, and training evaluation.

CHAPTER 2

LITERATURE REVIEW

Training

Training is typically defined as the systematic acquisition of knowledge, skills, or attitudes that results in improved performance in the work environment (e.g., Goldstein, 1993). This definition reveals two components of training: (a) the acquisition of knowledge, skills, or attitudes in the training environment and (b) the application of acquired knowledge, skills, and attitudes to the job environment.

The acquisition component of training occurs primarily in the training context and is often referred to as learning, "a relatively permanent change in knowledge or skill produced by experience" (Weiss, 1990, p. 172). The transfer component of training occurs primarily in the job context and is commonly referred to in the literature as the transfer of training (Baldwin & Ford, 1988; Goldstein, 1993). Within industrial and organizational psychology, transfer of training is usually defined as the degree to which trainees apply the knowledge, skills, and attitudes gained

in the training context to the job context (e.g., Ford & Kraiger, 1995; Tannenbaum & Yukl, 1992; Wexley & Latham, 1991).

While there are two distinct components of training, when a training evaluation is conducted, it usually focuses on the acquisition component. Yet, both training researchers and practitioners have generally acknowledged that the successful transfer of training to the job context is critical to the effectiveness of any training program (Baldwin & Ford, 1988; Beaudin, 1987; Campbell, Dunnette, Lawler, & Weick, 1970; Ford et al., 1992; Geber, 1995; Goldstein, 1993; Kelly, 1982; Kirkpatrick, 1976; Noe & Ford, 1992; Salas, Burgess, & Cannon-Bowers, 1995; Smith, Ford, Weissbein, & Gully, 1995; Tziner, Haccoun, & Kadish, 1991; Wexley & Latham, 1991; Yelon, 1992).

Baldwin and Ford (1988) suggested that two conditions are necessary for transfer to occur: (a) the learning must be generalized to the job context, and (b) the learning must be maintained over time on the job. Generalization concerns the extent to which the knowledge, skills, and attitudes acquired in the training context are exhibited in the job context, and are applied to different tasks or situations beyond those experienced in the training context (Adams, 1987; Baldwin & Ford, 1988; Ford & Kraiger, 1995). According to Ford and Kraiger, this condition goes

beyond mimicry by requiring trainees to exhibit behaviors identical or similar to those learned in training, in response to the similar yet non-identical stimuli presented in training. The second condition, maintenance, refers to the length of time that trained skills and behaviors continue to be used on the job (Baldwin & Ford, 1988; Ford & Kraiger, 1995). This condition requires that trainees not only demonstrate transfer some time after training, but continue to demonstrate transfer over time (Ford & Kraiger, 1995).

When training is applied on the job, there are three potential outcomes: (a) positive transfer, which occurs when trainees' learning in the training context results in their improved performance in the job context; (b) negative transfer, which occurs when trainees' learning in the training context results in their poorer performance in the job context; and (c) zero transfer, which occurs when trainees' learning in the training context results in no observable effect on their performance in the job context (Cascio, 1991; Russell & Wexley, 1988).

Training Evaluation

Training evaluation is defined as "the determination of the effectiveness of a training program" (Kirkpatrick, 1976, p. 18-2). Given

the acquisition and transfer components of training, training evaluation can be used to determine (a) the extent of learning in the training context, and ultimately, (b) the extent of transfer to the job context (Bunker & Cohen, 1977). Goldstein (1993) discussed the evaluation of these two components in terms of training goals. The goal of training validity asks if trainees learned anything during the training program, while the goal of transfer validity asks if the knowledge, skills, and attitudes learned in training resulted in improved performance on the job. Similarly, Bunker and Cohen (1977) asserted that the evaluation of these two components could be viewed as issues of internal and external validity, since internal (training) validity is a necessary, though insufficient, requirement for external (transfer) validity. As Goldstein (1993) stated, "unless the individual learns during training, the question of transfer to the actual job setting is meaningless" (p. 123).

If the fundamental purpose of training is to help trainees develop the knowledge, skills, or attitudes which, when applied on the job, will improve their performance, then the ultimate goal of training evaluation must be to determine the level of transfer (Tziner et al., 1991). In other words, a training program that results in learning in the training context is still inadequate if it fails to result in transfer on the job; or, to use a

medical analogy, the operation may have been a success, but the patient died (Leifer & Newstrom, 1980).

In their review of training evaluation research, Huczynski and Lewis (1980) found that early research focused primarily on the acquisition component of training and the training context in which it occurred.

> Course members were questioned on the content of the course, the teaching methods used in presenting the subject matter, the course design and so on. The focus of these numerous research studies was the training experience and, in the research designs, the course was treated as an end in itself. For the companies who paid for this training, however, the instruction was only a means to an end. In the majority of cases, that end could be defined as making the trainee more efficient or effective in the performance of his managerial duties and, in consequence, improving the performance of the organization as a whole. (p. 228)

This rather provincial view of training evaluation ignored the transfer component of training and the job context in which it occurs by assuming that if trainees learned what they were taught in training, they would automatically apply what they had learned on the job (Mosel, 1957). This assumption is reckless in that "the fact that one has learned something does not in any way guarantee that the knowledge will manifest itself in a change in performance" (Davis, 1979, p.125).

Subsequent research has demonstrated the flaw in this assumption by revealing the "transfer problem;" much of the training conducted in

organizations fails to transfer to the job context (Baldwin & Ford, 1988). For example, Huczynski and Lewis (1980) evaluated a course over a 9 year period only to find that, despite its popularity, the level of transfer was no higher than 10%. Furthermore, several training practitioners have estimated that only 10% of training expenditures actually result in a lasting behavioral change on the job (e.g., Kelly, 1982; Georgenson, 1982).

Concern over the transfer problem has spawned a great deal of research seeking to identify a host of individual and environmental variables which might explain why trainees are able to demonstrate learning at the conclusion of training, yet fail to transfer what they have learned to the job. Some of these variables include trainee motivation (Baldwin, Magjuka, & Loher, 1991; Clark, Dobbins, & Ladd, 1993; Facteau, Dobbins, Russell, Ladd, & Kudisch, 1995; Mathieu et al., 1992; Noe & Schmitt, 1986; Quinones 1995), trainee expectations (Cannon-Bowers, Salas, Tannenbaum, & Mathieu, 1995; Hicks & Klimoski, 1987), trainee self-efficacy (Cannon-Bowers et al., 1995; Ford et al., 1992; Gist, Stevens, & Bavetta, 1991; Quinones, 1995), trainee locus of control (Noe & Schmitt, 1986; Tziner et al., 1991), trainee ability (Baldwin et al., 1991; Ford et al., 1992; Tannenbaum, Mathieu, Salas, & Cannon-

Bowers, 1991), transfer climate (Clark et al., 1993; Tracey, Tannenbaum, & Kavanagh, 1995), and opportunity to perform (Ford et al., 1992). Other research has focused on transfer of training facilitation strategies such as goal setting and behavioral self-management (Gist, Bavetta, & Stevens, 1990; Gist et al., 1991; Wexley & Baldwin, 1986).

This problem of neglecting to evaluate trainee transfer of training might be solved if training researchers and practitioners adopt a more systematic view of the training evaluation process. Goldstein (1993) described training evaluation as a two-step process that (a) establishes measures of success, or criteria, for the acquisition and transfer components of training and (b) employs an appropriate research design to determine what changes have occurred in the training and job contexts.

Criteria

The evaluation of a training program generally requires the examination of several criteria (Cascio, 1991; Clement, 1982; Tannenbaum & Woods, 1992), as evaluators can be misled by a single criterion that fails to capture the diverse nature of the instructional process (Goldstein, 1993).

Cascio (1991) suggested that criteria could be distinguished by time, type, and source. With respect to time, criterion data could be gathered before, during, immediately after, or some time after the conclusion of training. Cascio recommended evaluating multiple time criteria because, for example, conclusions drawn from an analysis of trainee changes immediately after training may differ substantially from an analysis of trainee changes several months after training. With respect to type, criteria can be internal or external. While internal criteria are designed to assess trainee changes in the training context, external criteria are designed to assess trainee changes in the job context. Finally, criteria can be distinguished by source. While criterion data is usually obtained from the individual trainee, Cascio recommended obtaining data from additional sources including supervisors, peers, and subordinates.

Ideally, the use of multiple criteria refers not only to the selection of a single time, type, and source criterion, but to the selection of multiple time, type, and source criteria. For example, while Facteau et al.'s (1995) training evaluation study employed a self-report type criterion, they encouraged other researchers to obtain criterion data from additional sources such as supervisors, peers, and subordinates; to utilize measures

of learning, behavior, and organizational results; and to collect this data at multiple time intervals.

Kirkpatrick's (1976) hierarchical model of training evaluation proposed that training programs be evaluated with respect to four "levels," or types, of criteria: reaction, learning, behavior, and results criteria. Reaction criteria assess trainees' satisfaction with the content and administration of the training (Noe & Schmitt; 1986), and might consist of a simple questionnaire administered immediately following the program (Kirkpatrick, 1976). Learning criteria assess the knowledge, skills, and attitudes internalized by trainees during the training (Kirkpatrick, 1976), and are usually measured via a test or work sample administered to trainees at the end of a program (Ford & Kraiger, 1995). Behavioral criteria assess changes in trainees' training-related behaviors on the job (Mathieu et al., 1992), and typically consist of performance ratings, behavioral observations, or observational questionnaires obtained some time after trainees have returned to their jobs and have been given time to apply what they have learned (Ford & Kraiger, 1995; Salas et al., 1995). Finally, organizational results criteria assess the training program's contribution to the organization's objectives (e.g., decreased costs, increased profits, increased productivity, fewer accidents; Cascio,

1991). Results measures might consist of expenditure, savings, and output data obtained once trainees' new behaviors have been given time to impact organizational objectives (Salas et al., 1995).

Kirkpatrick's (1976) model has been widely accepted by training researchers (e.g., Facteau et al., 1995; Ferguson, 1968; Mathieu et al., 1992; Noe & Schmitt, 1986; Quinones, 1995), prompting several to recommend that training be evaluated by assessing some combination of its criteria (e.g., Salas et al., 1995; Tannenbaum & Woods, 1992). However, the model is not without its critics (e.g., Alliger & Janak, 1989; Clement, 1982; Kraiger, 1995; Newstrom, 1978) who have claimed that it makes at least three unverified assumptions. First, the model assumes that criteria are arranged in order of increasing value of the information to be gained by its evaluation (Alliger & Janak, 1989; Newstrom, 1978). Alliger and Janak refuted this assumption by reminding researchers that a criterion's value depends on the objectives of the particular training program. For example, researchers evaluating a company history training program might consider reaction and learning criteria to be of greater value than behavioral or results criteria.

Second, the model assumes that there exists a "positive manifold" such that all criterion levels are positively intercorrelated (Alliger & Janak,

1989; Clement, 1982; Newstrom, 1978). Yet, in a meta-analysis of training evaluation studies reporting two or more of the model's criteria, Alliger and his colleagues (Alliger & Janak, 1989; Alliger, Tannenbaum, & Bennett, 1995) examined the mean intercorrelations between reaction, learning, behavior, and results criteria only to find virtually no relationship between reaction criteria and other criteria, but slightly higher mean correlations between learning and behavior, behavior and results, and learning and results criteria.

Finally, the model assumes that subsequent criterion levels are caused by preceding criterion levels (Alliger & Janak, 1989; Clement, 1982; Hamblin, 1974; Mathieu et al., 1992; Noe, 1986). Thus, positive reactions to training are believed to promote learning, learning is believed to contribute to behavioral change on the job, and behavioral change is believed to lead to organizational results (Hamblin, 1974; Newstrom, 1978; Noe, 1986). Given the results of Alliger and Janak's meta-analysis, researchers do not widely support this assumption (e.g., Goldstein, 1993; Salas et al., 1995; Tannenbaum & Yukl, 1992). Even Kirkpatrick (1976) has acknowledged that a trainee's positive reaction to a training program neither assures that learning has taken place, nor does learning assure that a behavioral change will subsequently occur on the

job. At an anecdotal level, most trainees have probably experienced either liking (or not liking) a training program but not learning (or learning) the intended training objectives. Despite the criticism directed at the Kirkpatrick model (1976), it is still considered appropriate for evaluating training program effectiveness (Quinones, 1995).

Research Designs

The second step in Goldstein's (1993) evaluation process is the selection of an appropriate research design. Generally, training evaluation involves the use of either an experimental or quasi-experimental research design (Cascio, 1991). In their review of training and human resource management textbooks, Sackett and Mullen (1993) found that most texts presented the formal experimental design as the mechanism by which training programs should be evaluated. This design is characterized by the use of an experimental group which experiences the training, at least one control group which does not experience the training, the random selection and assignment of subjects to either the experimental or control groups, and the collection of pretraining and posttraining data (Arvey & Cole, 1989; Cascio, 1991; Goldstein, 1993; Tannenbaum & Woods, 1992). Researchers have described the experimental research design as

"standard" or "classic" (Arvey & Cole, 1989, p. 90), "traditional" (Bunker & Cohen, 1989, p. 527), and "adequate for most purposes" (Cascio, 1991, p. 397). The clear preference for the experimental design is largely due to its randomization component, which allows the researcher to conveniently rule out most alternate explanations for observed changes in trainees' knowledge, skills, and attitudes. This, in turn, permits the researcher to attribute the cause of these changes to the training program being evaluated.

Unfortunately, the rigorous nature of the experimental design rarely lends itself to the evaluation of training programs in organizational settings (Cascio, 1991; Goldstein, 1993; Haccoun & Hamtiaux, 1994; Sackett & Mullen, 1993). Researchers have long recognized the constraints which limit the feasibility of formal experimentation in organizations. Specifically, researchers are often unable to obtain pretests, control groups, and randomly assigned subjects from the organization sponsoring the training program. For example, the incorporation of pretests in the research design would require that the researcher be involved with the organization prior to the implementation of the training program. However, if the researcher intervenes during or

after the training program, the opportunity to collect pretest data will have passed.

Many organizations are also unable or unwilling to provide a control group. The inability to provide a control group might result when all employees are assigned to the experimental group because all are required to receive the training simultaneously. This situation is especially common in military training, where mission accomplishment and personnel safety might be endangered if soldiers, sailors, and airmen are placed in jobs without the appropriate training (Blaiwes, Puig, & Regan, 1973). An organization's unwillingness to provide a control group might stem from the perceived inequities that arise when employees in the experimental group are "awarded" desirable training while those in the control group are not. Also, those "selling" the training program may find themselves in the difficult position of having to convince the organization that their program is worth implementing while simultaneously arguing that the program needs to be evaluated using a control group.

Finally, organizations often hesitate to permit the random assignment of trainees to experimental and control groups. Not only does this practice disrupt the training of intact workgroups, but it might also be perceived by employees as an arbitrary system of rewarding employees

with training. Additionally, it may be difficult for the training researcher to convince the organization to suspend its practice of assigning trainees to training programs based on an individual's training needs.

Given that organizational environments are not conducive to the extensive use of experimental designs, researchers have suggested the exploration of alternative research designs (e.g., Cascio, 1991; Goldstein, 1980, 1993; Sackett & Mullen, 1993; Tannenbaum & Yukl, 1992). Quasi-experimental designs are a less rigorous alternative to experimental designs (Cook & Campbell, 1979). Unlike its more traditional counterpart, quasi-experimental designs do not rely on the random assignment of subjects to equate the experimental and control groups, and some do not even require the use of a control group (Cascio, 1991; Cook & Campbell, 1979; Cook , Campbell, & Perracchio, 1990; Tannenbaum & Woods, 1992). Without randomization to rule out most alternative explanations for changes in trainees' knowledge, behaviors, and attitudes, the quasi-experimental design requires that the researcher systematically evaluate and rule out each "threat" to internal validity (Cook & Campbell, 1979).

While the quasi-experimental design is more flexible than the experimental design, and therefore more conducive to real-world training evaluation in organizations, Sackett and Mullen (1993) found that its

divergence from the traditional experimental design had led several researchers to criticize it as inadequate.

> From the point of view of many textbooks on the topics of training and chapters on training in human resource management texts, a process is woefully inadequate as an evaluation if there is no pretest, and there is no control group at all, much less random assignment to experimental and control groups. Many texts would exhort students to strive for more rigorous evaluation that that described above. Some would go so far as to denounce the above process as useless. (p. 615)

Specifically, Sackett and Mullen (1993) objected to text written by Camp, Blanchard, and Huszczo (1986), which stated:

> We would suggest that no outcome evaluation at all is preferable to an evaluation without pretesting or some form of control group. Since evaluations without such precautions are without validity the company might as well save the money. (p. 167)

Sackett and Mullen's review prompted Haccoun and Hamtiaux (1994) to comment on the paradoxical situation whereby "those research designs which permit convincing training evaluations can rarely be implemented in organizations, while those designs that are practical for organizations are judged inadequate for evaluation research" (p. 594).

The purpose of an experiment is to eliminate alternative hypotheses which might account for the observed change in trainees' knowledge, skills, and attitudes (Cook & Campbell, 1979). While the experimental research design may be ideally suited to accomplish this task, the quasi-

experimental design is quite capable of eliminating some of these alternative hypotheses when use of an experimental design is impractical. Thus, the researcher's goal should be to create an experimental research design, but should this prove impossible given the constraints of the organization, the researcher should create the most rigorous design possible while recognizing its limitations (Cascio, 1991; Goldstein, 1993; Sackett & Mullen, 1993).

Due in part to the lack of availability of control groups in organizations, the one-group pretest-posttest design is one of the most common research designs used in organizational research (Cook et al., 1990). Yet, this design has been labeled "generally uninterpretable," as it is insufficient for permitting the researcher to make strong causal inferences about the success of a training program (Cook & Campbell, 1979; Cook et al., 1990). In response, Haccoun and Hamtiaux (1994) proposed the Internal Referencing Strategy (IRS) as a way of strengthening the design such that the researcher can estimate changes in trainees' knowledge, skills, and abilities, and make inferences regarding a training program's effectiveness.

Internal Referencing Strategy

Rationale and Methodology

The IRS is essentially a one-group pretest-posttest design in which the pretests and posttests include items that are both relevant and irrelevant to the training program. The rationale for this is straightforward. In a traditional one-group pretest-posttest design, higher posttest scores relative to pretest scores is a necessary though insufficient indication of training effectiveness. An observed difference is insufficient because the design is incapable of ruling out a number of alternative explanations, or threats to internal validity, that might be responsible for these differences. According to Cook et al. (1990), "threats to internal validity compromise the inferences about whether the relationship observed between two variables would have occurred even without the treatment under analysis" (p. 500). Thus, in the case of training evaluation, the observed differences between pretest and posttest might not be due to the effectiveness of the training program at all, but some extraneous factor (e.g., maturation of subjects). By incorporating pretest and posttest items that are both relevant and irrelevant to the training program, the IRS approach improves upon the one-group pretest-

posttest design by providing a standard of comparison by which the researcher can eliminate several threats to internal validity.

In the IRS approach's evaluation of training effectiveness, the researcher examines the patterns of changes in the relevant items relative to the patterns of changes in the irrelevant items. In order to be judged a successful training program, planned changes among relevant items must be greater than unplanned changes among irrelevant items. In short, the researcher can infer training effectiveness when significantly greater, positive changes are found on relevant items as compared to irrelevant items.

Irrelevant Items

When applying the IRS approach, great care must be exercised when selecting irrelevant items. Irrelevant items could be drawn from training objectives or training content that logically could have been included in the training program being evaluated, but for whatever reason (e.g., cost or time constraints, needs of the customer) were not (Haccoun & Hamtiaux, 1994). The method of selecting irrelevant items is critical, as "plausible alternative interpretations have to be specified in terms of construct-related differences in maturation rates, local history, or

instrumentation" (Cook et al., 1990, p. 547). For example, it would be relatively useless to demonstrate that a data analysis training program was associated with greater learning on test items relating to data analysis, but not associated with greater learning on test items relating to car repair. A more appropriate choice of irrelevant domain might be business statistics, as the irrelevant domain must be related to the relevant domain, such that it would be expected that observed differences in learning and behavior among both the relevant and irrelevant items would be equally affected by alternative explanations such as history, maturation, instrumentation, or testing.

Validity Issues

Like any other research design, the utility of the IRS approach depends upon its ability to eliminate alternative explanations, or threats to internal validity, for the observed results of the training evaluation. Cook and Campbell (1979) described thirteen threats to internal validity: history, maturation, testing, instrumentation, ambiguity about the direction of causal influence, statistical regression, selection, interactions with selection, diffusion or imitation of treatments, compensatory equalization of treatments, compensatory rivalry by respondents receiving

less desirable treatments, and resentful demoralization of respondents receiving less desirable treatments. In a traditional experimental research design, the random selection and assignment of trainees to experimental and control groups serves to eliminate many of these thirteen threats to internal validity. However, quasi-experimental designs, including the IRS approach, lack this randomization component and therefore depend upon the researcher to systematically investigate and eliminate each threat given the nature of the individual research design.

In evaluating the IRS approach's capacity to account for each of the thirteen threats, the latter eight can be eliminated from consideration as they are "between-group" threats (Haccoun & Hamtiaux, 1994). Between-group threats arise from a lack of comparability between the experimental and control group which might "bias the composition of the experimental and comparison groups" and/or "contaminate the comparison condition by including in it parameters that should have been reserved exclusively for the experimental group" (Haccoun & Hamtiaux, 1994, p. 596). As the IRS approach does not require a control group, between-group threats are not relevant.

However, the researcher employing the IRS approach should be concerned with the "within-group" threats: history, maturation, testing,

and instrumentation (Cook et al., 1990). History is a threat when the relationship between the training program and the pretest and posttest differences might be due to an external event that occurred between pretests and posttests, but was not part of the training program (Cook et al., 1990). The IRS approach infers training effectiveness when pretest and posttest differences are greater for relevant items than for irrelevant items. Therefore, for history to threaten the inferences resulting from this approach, it would have to increase the pretest and posttest differences on the relevant items but not the irrelevant items, contributing to a Type I error and the false conclusion of training effectiveness. According to Haccoun and Hamtiaux (1994), because both the relevant and the irrelevant items are extracted from the same general content domain, it is unlikely that history would selectively affect the relevant items but not the irrelevant items.

Maturation is a threat when the relationship between the training program and the pretest and posttest differences might be due to trainees growing wiser or more experienced between pretests and posttests (Cook et al., 1990). Like history, the presence of a maturation threat could lead to a Type I error. While, Haccoun and Hamtiaux (1994) do not expect

maturation to affect the relevant items but not the irrelevant items, they do raise the possibility of a maturation-history interaction.

> For example, a course on DOS may heighten a person's awareness and sensitivity to other information about computer functioning, including, for example WINDOWS. Posttest differences between (irrelevant) WINDOWS knowledge and target (DOS) knowledge would then be smaller, arguing falsely against inferences of training effectiveness. The impact of maturation-history interaction may make the IRS approach vulnerable to a Type II error. (p. 596)

Testing is a threat when the relationship between the training program and the pretest and posttest differences might be due to trainees' familiarity and experience with the pretests contributing to their improved performance on the posttests (Cook & Campbell, 1979). Once again, for testing to threaten the IRS approach, the testing confound would have to selectively affect the relevant items but not the irrelevant items. As trainees will be exposed to the irrelevant items as often as their relevant counterparts, Haccoun and Hamtiaux (1994) conclude that the potential for Type I error is unlikely. However, they do concede the possibility of a Type II error if instrumentation problems aggravate testing effects such that "a constant increase from pretest scores that are initially high may effectively reduce differences between relevant and irrelevant posttest items" (p. 597). This threat can be mitigated through the use of

parallel forms of pretests and posttests and by selecting items with relatively high difficulty levels to prevent ceiling effects.

The final threat, ambiguity about the direction of causal influence, can also be eliminated as this threat is not salient in studies in which the order of temporal precedence is clear (Cook et al., 1990). This threat is most salient in many non-experimental studies, especially those requiring a correlational analysis, where it is not clear if variable A caused variable B, or if variable B caused variable A. However, in training evaluation research, including evaluations using the IRS approach, the training program precedes the assessment of trainee learning and transfer. Therefore, if training effectiveness is inferred, and all other threats to internal validity are controlled, this ordering extinguishes any ambiguity concerning the causal relationship between the training program and the observed trainee learning and transfer.

Overall, Haccoun and Hamtiaux (1994) concluded that the IRS approach appeared to be more vulnerable to Type II error than to Type I error: while the approach has the potential to identify effective training programs, it may at times fail to identify effective training programs. Accordingly, the IRS approach should be used with caution when the

false rejection of an effective training program is a particularly undesirable error.

IRS Research

The IRS approach is not new; Cook and Campbell (1979) and later Cook et al. (1990), referred to it as the Nonequivalent Dependent Variables Design and suggested it might be incorporated into other research designs as a means of strengthening the inferences drawn from them. Similarly, Trochim (1985) discussed the merits of the design and also suggested its use in improving methods of evaluating social programs. Trulson (1986) used the Nonequivalent Dependent Variables Design to evaluate the effect of Tae Kwon Do martial arts training on the behavior of delinquent boys. He found that Tae Kwon Do reduced delinquent boys' scores on those (relevant) mental health constructs that distinguished delinquent boys from non-delinquent boys, but did not reduce scores on those (irrelevant) constructs that did not distinguish between the two groups. Additionally, Haccoun and Hamtiaux (1994) used the IRS approach in a training evaluation study.

The purpose of Haccoun and Hamtiaux's (1994) study was to validate the IRS approach by demonstrating that it permitted conclusions

regarding the effectiveness of a training program that were similar to those obtained from a more complex design. To achieve this end, they used the IRS approach to evaluate the knowledge acquisition of 42 mid-level managers of a large university who voluntarily attended a human resource management course. A group of 24 managers who did not attend the course served as a comparison group for validation purposes. The results of the evaluation demonstrated that the patterns of observed differences between pretest and posttest scores on relevant and irrelevant items for the experimental group alone (the IRS approach) permitted similar inferences regarding the effectiveness of the training program as compared to an evaluation of the patterns of observed differences between pretest and posttest scores on only relevant items for both the experimental and comparison groups (a more complex two-group pretest-posttest approach).

Their results also demonstrated the IRS approach's potential to strengthen the one-group pretest-posttest design. In the latter design, all pretest and posttest differences on relevant items could be considered evidence of an effective training program, leading the researcher to conclude that the training program was effective. However, the incorporation of the IRS approach might demonstrate that despite the

observed differences on relevant items, there was a corresponding difference in trainees' performance on the irrelevant items, prompting the researcher to conclude that the training program may not have been as effective as first thought.

Present Study

The current study has two purposes: (a) to conceptually replicate Haccoun and Hamtiaux's (1994) evaluation of trainee learning in the training context using the IRS approach and (b) to extend their study to the application of the IRS approach to the evaluation of trainee transfer of training to the job context.

Replicating the Learning Evaluation

In the original study, Haccoun and Hamtiaux (1994) used the IRS approach to infer the effectiveness of the acquisition, or learning, component of a human resource management training program. This was accomplished by an examination of trainees' performance on a learning measure consisting of pretraining and posttraining tests of declarative knowledge. This study conceptually replicates its predecessor in a corporate setting, involving a greater number of subjects, and using a

complex managerial training course. Additionally, in contrast to Haccoun and Hamtiaux study, this course is computer-administered and mandatory for all potential subjects.

Extending To a Transfer Evaluation

While Haccoun and Hamtiaux's (1994) original study applied the IRS approach to the evaluation of a training program with respect to learning, this study extends their research by applying the IRS approach to training evaluation with respect to the application, or transfer, component of training. Whereas Haccoun and Hamtiaux inferred learning based on pretest-posttest differences on tests of declarative knowledge, this study infers transfer based on pretest-posttest differences on a behavioral questionnaire. The effects observed in the learning evaluation are predicted to generalize to the transfer evaluation. Thus, two hypotheses are offered to replicate and extend Haccoun and Hamtiaux's support of the IRS approach as a research design for use in training evaluation.

Hypothesis 1: For trained subjects, there will be an interaction between time (pretest-posttest) and item relevancy (relevant-

irrelevant) such that pretest-posttest differences on the learning measure will be greater for relevant rather than irrelevant knowledge items.

Hypothesis 2: For trained subjects, there will be an interaction between time (pretest-posttest) and item relevancy (relevant-irrelevant) such that pretest-posttest differences on the transfer measure will be greater for relevant rather than irrelevant behavioral items.

CHAPTER 3

METHOD

Participants

The potential population of participants for this study consisted of 465 front-line (i.e., middle and upper-middle level) managers employed by a major regional telecommunications company. Although participants were stationed at different offices located throughout the company's multi-state region, all were associated with a revenue-producing division. Participant sex and age data were not provided by the organization.

Participants were selected by virtue of their job position within the company: all front-line managers were required to experience a managerial training program and participate in its evaluation. Due to the company's requirement that all managers undergo training simultaneously, a suitable control group was unavailable, and the random selection and assignment of subjects to the training group was not possible. Of the 465 potential participants in the study, only 39% provided usable pretest

and posttest scores on both the learning and transfer measures, yielding a final subject pool of 182.

Training Course

Marketplace Awareness, the course evaluated in this study, was the first in a series of six managerial training courses administered to front-line managers. Its content covered three general topics: (a) the history of the telecommunications industry, (b) its current competitive environment, and (c) the company's marketing strategy. While the course's declarative knowledge-based content facilitated the creation of relevant knowledge items for the learning measure, it hindered the creation of relevant behavioral items for the transfer measure. However, the course was considered suitable for this study's learning and transfer evaluation, as even declarative knowledge has behavioral implications for the job environment.

Due to the geographical diversity of the managers' offices, the course was administered via computer using the company's intranet service. The training course began with a videoconference course overview, followed with a 2 week training window during which trainees were required to complete the self-paced intranet course, and concluded

with a videoconference course summary. Upon course completion, managers were expected to (a) understand the telecommunications industry and marketplace; (b) understand key industry terms, concepts, and business issues; and (c) apply industry knowledge to their daily, management activities in support of the company's overall marketing strategy.

Measures

Learning Measure

Subject learning was evaluated using a 20-item test of declarative knowledge. The total number of items included in the learning measure was constrained by the time available for subject testing. Items were derived from an analysis of the contents of the training program, and employed a multiple choice format offering five possible responses. Alternate forms of a sample knowledge item are presented in Figure 3.1.

FIGURE 3.1

Sample Knowledge and Behavioral Items

Sample Knowledge Item (Form A)

A signal that uses 1's and 0's to represent voice, video, or data and can be compressed very efficiently is called a(n):

a. broadband signal.
b. analog signal.
c. digital signal.
d. centrex signal.

Sample Knowledge Item (Form B)

Unlike _____ signals, _____ signals use discrete values of 1's and 0's to represent voice, video, or data.

a. broadband; analog
b. analog; digital
c. digital; analog
d. digital; broadband

Sample Behavioral Item

Within the past two weeks, I have discussed strategies for marketing new packages or bundles of services.

1. Never demonstrated this behavior.
2. Demonstrated this behavior a few times (about 1-3 times).
3. Demonstrated this behavior occasionally (about 4-7 times).
4. Demonstrated this behavior frequently.
5. Not applicable.

Following Haccoun and Hamtiaux's (1994) example, 75% of the knowledge items (i.e., 15 items) were written to be relevant to the content of the training course, and 25% of the items (i.e., 5 items) were written to be irrelevant to the course. Although Haccoun and Hamtiaux did not specifically prescribe a 4:1 ratio of relevant to irrelevant items when implementing the IRS approach, this technique was adopted since one purpose of this study was to conceptually replicate their study using a different training program and subject pool.

Transfer Measure

Subject transfer of training was evaluated using a 15-item questionnaire of on-the-job behaviors. The questionnaire requested that subjects indicate the degree to which they had demonstrated a list of behaviors within the past 2 weeks. Each behavioral item was derived from an analysis of the contents of the training program and offered five possible responses: 1 (never demonstrated this behavior), 2 (demonstrated this behavior a few times [about 1-3 times]), 3 (demonstrated this behavior occasionally [about 4-7 times]), 4 (demonstrated this behavior frequently), and 5 (not applicable). A sample behavioral item is shown in Figure 3.1.

The length of the behavioral questionnaire was constrained by the time available for subject testing as well as the number of behaviors that trained subjects could reasonably be expected to demonstrate in the 2 weeks between the training course and the questionnaire's administration. This delay was imposed by the organization in order to conclude all Marketplace Awareness-related activities prior to the initiation of a subsequent course in the six course managerial training program.

During questionnaire development, an attempt was made to maintain a 4:1 ratio of relevant to irrelevant items (i.e., 11 or 73% relevant items and 4 or 27% irrelevant items) in congruence with the learning measure. However, the training material possessed a strong declarative knowledge orientation which prevented the writing of an adequate eleventh relevant behavioral item. Therefore, the final behavioral questionnaire consisted of 10 relevant and 5 irrelevant behavioral items, establishing a 3:1 ratio of relevant to irrelevant items (i.e., 67% relevant items and 33% irrelevant items).

Item Relevance

Relevant items for both the learning and transfer measures were derived from course material that was clearly linked to an established

Marketplace Awareness learning objective. Conversely, irrelevant items were not derived from Marketplace Awareness course material, as irrelevant items should be drawn from training objectives or training content that logically could have been included in the training program being evaluated, but for whatever reason were not (Haccoun & Hamtiaux, 1994). Thus, irrelevant items were derived from material linked to the established learning objectives of one or more follow-on courses within the six course series.

This method of irrelevant item writing was selected due to the relationship between the courses. Specifically, these six courses were the product of a single training needs assessment of company front-line managers, and were designed to flow from one course to the next, with each subsequent course building on the knowledge, skills, and abilities provided by its predecessor. Therefore, the six courses are linked such that training material from subsequent courses logically could have been included in Marketplace Awareness. However, as this would have created a unreasonably lengthy course, instructional designers elected to separate the training material into six manageable courses. Nevertheless, it is this link between courses that permitted irrelevant knowledge and behavioral items for the learning and transfer measures to be created from

from course material included in Data Analysis and Front-line Leadership, two subsequent courses in the six course series.

Measurement Validity and Reliability

The process of writing items linked to course training objectives not only served to create relevant items, but also promoted the content validity of the two measures. All 35 learning and transfer items were judged to be appropriately relevant or irrelevant by three Ph.D.-level training practitioners.

The reliability of the transfer measure's relevant behavioral item "sub-test" was assessed using Cronbach's (1951) coefficient alpha in order to provide an estimate of its internal consistency. The 10-item relevant behavior sub-test demonstrated acceptable internal consistency both at pretest (.78) and posttest (.77). The reliability of the relevant knowledge item sub-test could not be estimated due a computer error which resulted in the loss of item-level knowledge data.

Procedure

Learning Measure

Both the learning measure's knowledge pretests and posttests were administered to subjects via computer. Each subject completed a knowledge pretest on the day of the videoconference course overview and was then allotted a 2 week training window in which to complete the self-paced training course. Knowledge posttests were administered to subjects as they completed the course, regardless of when course completion was achieved relative to other subjects.

To discourage subjects from engaging in "group test-taking," all knowledge items were presented in random order both at pretest and posttest. Additionally, to mitigate the effects of practice, subjects received at posttest an alternate version of each knowledge item received at pretest. This was accomplished by writing two versions of each relevant and irrelevant knowledge item. As shown in Figure 3.1, each item counterpart was written to closely parallel the original question. When administering the knowledge tests, the computer program randomly selected a "version A item" or a "version B item" for presentation at pretest, then selected its alternate version for presentation at posttest. This method of item administration resulted in each subject receiving a

different combination of version A and version B knowledge items at pretest and posttest. Consequently, the relevant knowledge sub-test reliability estimates presented previously are based on the assumption that all possible pretest-posttest combinations of version A and version B knowledge items are equivalent. However, as both coefficients alpha are relatively high (i.e., .78 at pretest and .77 at posttest), it is expected that reliability estimates based on an all-version A-item pretest and all-version B-item posttest would have been higher.

Transfer Measure

Subjects completed the transfer measure's behavioral pretests immediately following their knowledge pretests on the day of the videoconference course overview. However, their behavioral posttests were administered 2 weeks after the close of the training window. This delay was intended to provide time for the managers to develop and demonstrate the skills they had acquired during training. Organizational constraints precluded a longer delay. Neither randomized item presentation nor alternate item versions were used in implementing the transfer measure.

Analysis

Data Preparation

Prior to the analysis, three pairs of scores were calculated for each subject: (a) a pretest and posttest relevant knowledge learning score, (b) a pretest and posttest relevant behavior transfer score, and (c) a pretest and posttest irrelevant behavior transfer score. In congruence with the two pairs of transfer scores, a second pair of learning scores (i.e., a pretest and posttest irrelevant knowledge learning score) should also have been calculated for each subject. However, a computer error, which resulted in the loss of irrelevant knowledge item-level data, prevented this calculation. Consequently, the learning results predicted by the study's first hypothesis could not be tested as planned.

Subjects' pretest and posttest relevant knowledge learning scores reflect the percentage of items they answered correctly on their pretest and posttest relevant knowledge item sub-tests. Each subject's four behavioral transfer scores consist of an average behavioral item rating based on only those items with valid responses.

Analytic Strategy

For hypothesis one, a paired samples *t* test was selected to analyze subjects' pretest and posttest relevant knowledge learning scores. While a repeated measures analysis of variance (ANOVA) had been planned, the loss of pretest and posttest irrelevant knowledge item-level data due to computer error necessitated the change in strategy. However, in keeping with the conceptual replication of Haccoun and Hamtiaux's (1994) IRS study, hypothesis two was analyzed using a repeated measures ANOVA consisting of two within-subject factors: (a) a time (pretest-posttest) factor, and (b) an item relevance (relevant-irrelevant) factor. Although an *F* test was available for the main effects of time and item relevancy, the interaction between the two within-subject factors was the focus of this analysis, as the presence of a significant interaction would indicate trainee transfer of training.

CHAPTER 4

RESULTS

In the evaluation of Marketplace Awareness, two hypotheses were offered. An investigation of hypothesis one would establish whether or not trainee learning had occurred and an investigation of hypothesis two would determine whether or not trainee transfer of training had occurred.

Learning Evaluation

Hypothesis one, which served as an evaluation of Marketplace Awareness with respect to trainee learning, predicted an interaction between time (pretest-posttest) and item relevancy (relevant-irrelevant) such that pretest-posttest differences on the learning measure would be greater for relevant rather than irrelevant knowledge items. Due to the loss of irrelevant knowledge item-level data, this hypothesis could not be investigated as planned (i.e., using a repeated measures ANOVA). Instead, the pretest and posttest relevant knowledge learning scores were analyzed using a paired samples *t* test. Table 4.1 presents the means and standard deviations associated with the pretest and posttest relevant

knowledge sub-tests. With an alpha level of .05, a paired samples *t* test revealed a significant difference between pretest and posttest relevant knowledge learning scores [$t(181) = 9.87$, $p < .001$; see FIGURE 4.1]. Thus, this result provided some evidence that trainees had acquired relevant knowledge during Marketplace Awareness.

FIGURE 4.1

Relevant Knowledge Item Sub-Test Scores at Pretest and Posttest

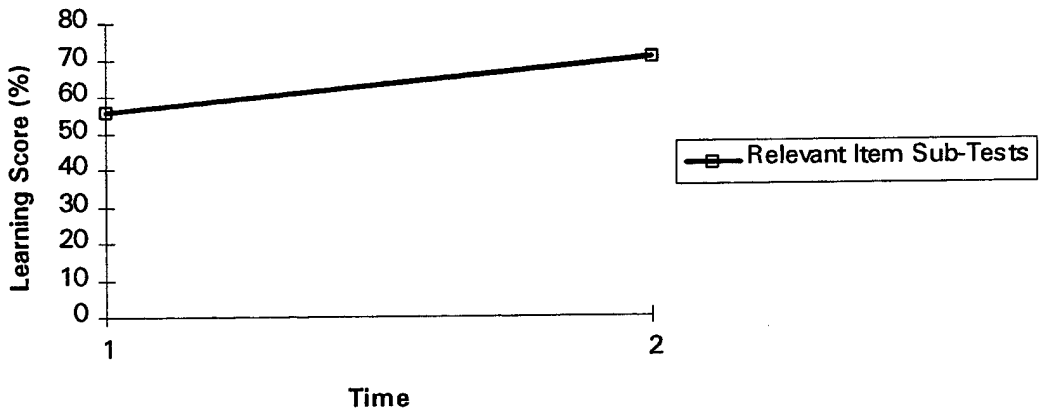


TABLE 4.1

Descriptive Statistics of Sub-Test Combinations

Sub-Test	M	SD	n
Relevant Knowledge			
Pretest	55.72	13.41	182
Posttest	70.07	14.32	182
Relevant Behavior[1]			
Pretest	2.75	0.30	182
Posttest	2.93	0.29	181
Irrelevant Behavior[1]			
Pretest	2.51	0.64	181
Posttest	2.64	0.63	180
Relevant Behavior[2]			
Pretest	2.50	0.61	182
Posttest	2.73	0.61	181
Irrelevant Behavior[2]			
Pretest	1.79	0.70	175
Posttest	1.91	0.71	170

[1]Based on the initial analysis and included all behavioral items.

[2]Based on the subsequent analysis and excluded those behavioral items exhibiting ceiling effects.

Transfer Evaluation

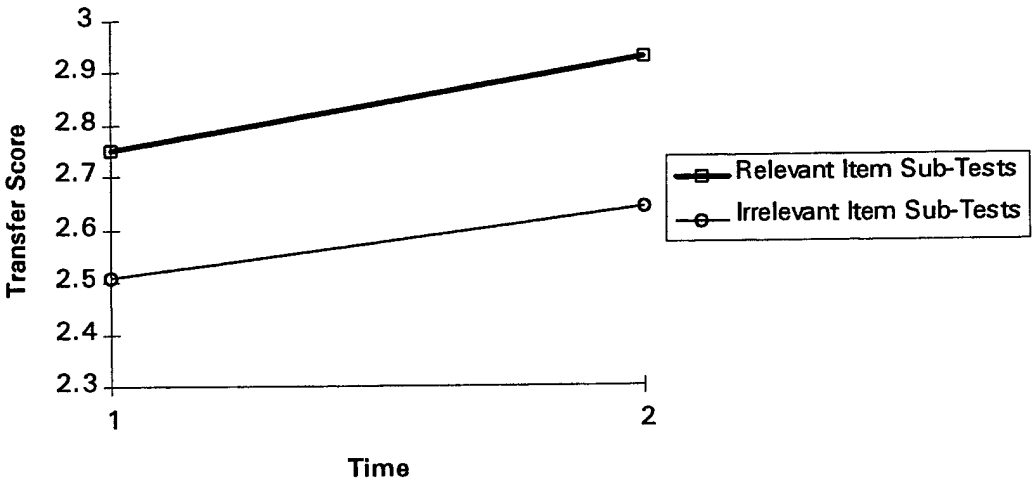
As an evaluation of Marketplace Awareness with respect to trainee transfer of training, hypothesis two predicted an interaction between time (pretest-posttest) and item relevancy (relevant-irrelevant) such that pretest-posttest differences on the transfer measure would be greater for relevant rather than irrelevant behavioral items. Table 4.1 presents the means and standard deviations associated with the four pretest-posttest and relevant-irrelevant behavioral item sub-test combinations. With an alpha level of .05, a repeated measures ANOVA revealed significant main effects for time [$E(1,178) = 57.41$, $p < .001$, $Eta^2 = .24$] and item relevance [$E(1,178) = 11.78$, $p = .001$, $Eta^2 = .06$], but no significant time-by-item relevance interaction [$E(1,187) = 2.03$, $p = .279$, $Eta^2 = .01$]. The significant main effects and non-significant interaction for hypothesis two are illustrated in Figure 4.2.

The significant main effect for time revealed that trainees engaged in significantly more behaviors after training than before training, regardless of their degree of relevancy (or irrelevancy). The significant main effect for item relevancy indicated that trainees engaged in significantly more relevant behaviors than irrelevant behaviors, regardless

of their test occasion. However, the absence of a significant interaction revealed that the number of relevant behaviors trainees engaged in at posttest relative to pretest was not significantly greater than the number of irrelevant behaviors they engaged in at posttest relative to pretest. Consequently, hypothesis two, which predicted trainee transfer of training, was not supported.

FIGURE 4.2

Initial Analysis of Relevant and Irrelevant Behavioral Item Sub-Test Scores at Pretest and Posttest



However, it is possible that this conclusion is false, as Haccoun and Hamtiaux (1994) stated that the IRS approach appeared to be "more vulnerable to a Type II error than a Type I error" (p. 597). Such an error might occur in the presence of one or more threats to the study's internal validity. Specifically, Haccoun and Hamtiaux had warned against an instrumentation threat resulting from ceiling effects.

In order to avoid ceiling effects in their investigation, Haccoun and Hamtiaux (1994) pilot tested their relevant and irrelevant items, retaining only those items with relatively high difficulty levels (about 50%). Unfortunately, time constraints precluded a similar pilot test prior to this evaluation. An examination of pretest relevant and irrelevant behavioral item means (shown in Table 4.2) revealed two relevant and two irrelevant items exhibiting ceiling effects (i.e., pretest item means greater than or equal to the rating scale midpoint of 3.00 accompanied by small standard deviations). Despite its large standard deviation, a third relevant behavioral item (i.e., item 6) was determined to exhibit ceiling effects based on a combination of its high pretest mean and high negative skew. Thus, the inclusion of these items in the initial analysis may have masked any trainee transfer of training.

TABLE 4.2

Descriptive Statistics for Relevant and Irrelevant Behavioral Items at Pretest

Behavioral Items	M	SD	Skew	95% CI	n
Relevant					
Item 1[1]	3.71	0.51	-1.59	(3.63, 3.79)	175
Item 2	2.92	0.84	-0.16	(2.80, 3.04)	168
Item 3	2.02	1.04	0.72	(1.82, 2.22)	115
Item 4[1]	3.63	0.66	-1.54	(3.53, 3.73)	172
Item 5	2.97	0.85	-0.28	(2.85, 3.09)	178
Item 6[1]	3.06	1.14	-0.75	(2.86, 3.26)	131
Item 7	2.04	0.95	-0.61	(1.88, 2.20)	156
Item 8	1.97	0.99	0.75	(1.81, 2.13)	147
Item 9	1.84	1.00	1.01	(1.66, 2.02)	135
Item 10	2.58	0.93	0.64	(2.44, 2.72)	171
Irrelevant					
Item 1	1.76	0.85	1.01	(1.62, 1.90)	153
Item 2	2.12	0.98	0.59	(1.96, 2.28)	150
Item 3	1.34	0.63	1.86	(1.24, 1.44)	139
Item 4[1]	3.52	0.77	-1.38	(3.42, 3.66)	168
Item 5[1]	3.40	0.84	-1.11	(3.28, 3.52)	167

[1]Items exhibiting ceiling effects.

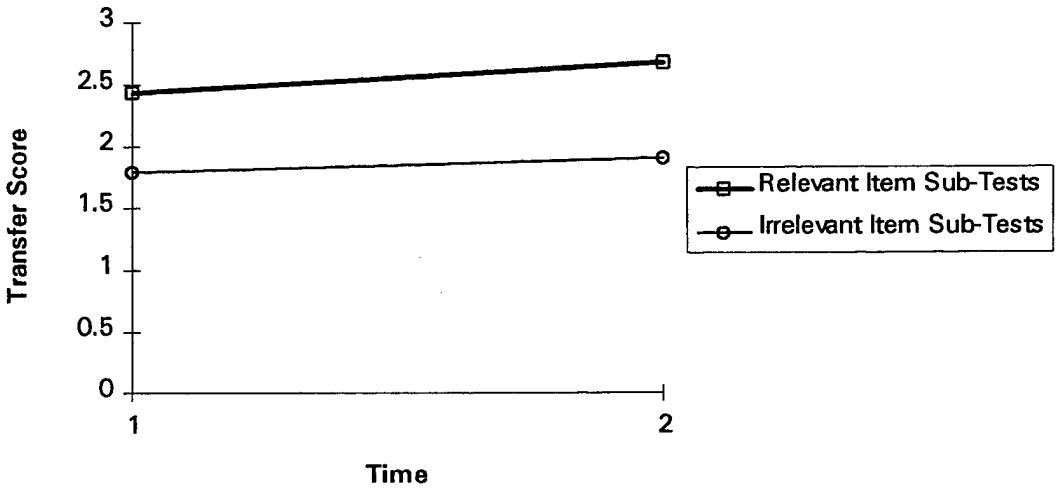
To test this explanation, a second repeated measures ANOVA was conducted after excluding from the analysis the five behavioral items exhibiting ceiling effects. Once again, Table 4.1 presents the means and standard deviations associated with the four new pretest-posttest and relevant-irrelevant behavioral item sub-test combinations. With an alpha level of .05, the second repeated measures ANOVA revealed significant main effects for time [$E(1,164) = 264.75$, $p < .001$, $Eta^2 = .62$] and item relevance [$E(1,164) = 15.69$, $p < .001$, $Eta^2 = .09$], as well as a significant time-by-item relevance interaction [$E(1,164) = 4.64$, $p = .03$, $Eta^2 = .03$]. The significant main effects and interaction are illustrated in Figure 4.3.

As in the initial analysis (shown in Figure 4.2), Figure 4.3 depicts significant main effects for time and item relevance. However, unlike the initial analysis, the subsequent analysis revealed a significant time-by-item relevancy interaction: the number of relevant behaviors trainees engaged in at posttest relative to pretest was significantly greater than the number of irrelevant behaviors they engaged in at posttest relative to pretest. Thus, by excluding those behavioral items exhibiting ceiling effects,

trainee transfer of training was "unmasked," revealing support for hypothesis two.

FIGURE 4.3

Subsequent Analysis of Relevant and Irrelevant Behavioral Item Sub-Test Scores at Pretest and Posttest



CHAPTER 5

DISCUSSION

The purpose of this study was to (a) conceptually replicate Haccoun and Hamtiaux's (1994) use of the IRS approach to evaluate trainee learning in the training context and (b) to extend their study to the application of the IRS approach to the evaluation of trainee transfer of training to the job context. The successful application of the IRS approach in both investigations would provide training practitioners with a rigorous research methodology for use in training evaluations in field settings.

Summary of Results

Learning Evaluation

The first hypothesis predicted an interaction between time (pretest-posttest) and item relevancy (relevant-irrelevant) such that pretest-posttest differences on the learning measure would be greater for relevant rather than irrelevant knowledge items. Due to the unfortunate loss of irrelevant knowledge item-level data, the planned repeated measures

ANOVA could not be performed. However, a paired samples *t* test of the remaining data revealed a significant difference between pretest and posttest relevant knowledge learning scores. While this result provided some evidence of the effectiveness of Marketplace Awareness with respect to trainee learning, it could not conclusively attribute the observed result to training effectiveness alone. Without a standard of comparison from which to interpret the results (i.e., a control group's relevant knowledge learning scores or the treatment group's irrelevant knowledge learning scores), it is possible that the observed results may have been due to one or more threats to internal validity (e.g., history or maturation).

Transfer Evaluation

The second hypothesis predicted an interaction between time (pretest-posttest) and item relevancy (relevant-irrelevant) such that pretest-posttest differences on the transfer measure would be greater for relevant rather than irrelevant behavioral items. Despite the presence of significant main effects for the time and item relevancy within-subject factors, the initial repeated measures ANOVA failed to support this hypothesis by revealing a non-significant time-by-item relevancy interaction. This pattern of results indicated that transfer of training did

not occur among managers. However, as previously stated, the IRS approach is vulnerable to a Type II error in the presence of ceiling effects or a maturation-history interaction. For this reason, a subsequent repeated measures ANOVA was performed, after removing those items exhibiting ceiling effects. This analysis revealed a significant time-by-item interaction indicating trainee transfer of training did occur among managers.

Due to the manner in which the transfer evaluation was implemented, a maturation-history interaction might also have contributed to a Type II error. This is because irrelevant behavioral items were drawn from the training content of subsequent courses in the six course managerial training program. Prior to Marketplace Awareness, pre-training publicity present within the organization (i.e., an overview of the content of the six courses) may have sensitized subjects to the irrelevant domain (i.e., the training content from subsequent courses), resulting in subjects' improved performance on the irrelevant behavioral item sub-tests. This could have masked the level of transfer among relevant behaviors relative to irrelevant behaviors, which in turn could have contributed to the false conclusion of trainee transfer of training ineffectiveness.

Implications

Learning Evaluation

Due to a loss of data, this study was unable to apply the IRS approach to the evaluation of trainee learning. However, this unfortunate occurrence does not prompt the conclusion that the approach should not be used in future training evaluations. On the contrary, the IRS approach should be applied whenever possible, and especially in those cases in which a control group is unavailable, as Haccoun and Hamtiaux's (1994) investigation had previously demonstrated the IRS approach's ability to infer training effectiveness with respect to trainee learning. Given the multitude of organizational constraints that reduce both the quantity and quality of training evaluations conducted in field settings, the IRS approach represents a boon to those training practitioners in need of a rigorous, flexible, and easily implemented research methodology.

Transfer Evaluation

The initial results of this study failed to support the application of the IRS approach to the evaluation of trainee transfer of training. However, the success of the subsequent investigation, which excluded data exhibiting ceiling effects, demonstrated the approach's potential for

use in transfer of training evaluations. Thus, while the approach may be useful for assessing trainee transfer of training, it also appears that Haccoun and Hamtiaux's (1994) warning that the approach is vulnerable to Type II errors is well founded. Clearly, the results of the initial and subsequent repeated measures ANOVAs illustrate the significant threat ceiling effects pose to the implementation of the IRS approach. If the IRS approach is to succeed in transfer of training evaluations, the evaluator must guard against the dual threats of a instrumentation due to ceiling effects and a maturation-history interaction.

Limitations

Learning Evaluation

The conclusions that can be drawn from the learning evaluation are limited by the loss of irrelevant knowledge item-level data which would have served as a standard of comparison from which to interpret the trainee learning results. Although the paired samples *t* test of relevant knowledge learning scores revealed a significant pretest-posttest difference, trainee learning cannot be concluded with confidence without the benefit of the lost pretest and posttest irrelevant knowledge learning scores and their capability to disqualify numerous threats to the test's

validity. Had the organization been able to provide a suitable control group, the learning evaluation could have recovered from this loss of data by employing a more traditional two-group pretest-posttest design involving an analysis of relevant knowledge learning scores to infer trainee learning.

Based on trainee comments regarding the learning evaluation, trainee reactance may have affected their performance on the posttest knowledge sub-tests. Specifically, trainees would often became frustrated when they encountered an irrelevant knowledge item at posttest. Having just completed the training program, many would struggle to recall the appropriate training material in order to select the correct response. However, as irrelevant items were derived from training material obtained from a subsequent course, this effort was destined to fail. Consequently, frustration-induced reactance may have affected subject performance on the posttest learning measure.

While the computer program administering the training program had explained the presence and role of irrelevant items in the evaluation, many subjects apparently chose to ignore these directions (i.e., press any key) and commence the evaluation. In future studies, this potential problem

might be remedied by using a more salient format to emphasize the irrelevant item aspect of the evaluation.

Transfer Evaluation

Throughout this study, the transfer evaluation was plagued by time constraints which ultimately compromised the methodological integrity and implementation of the IRS approach. These compromises included (a) the choice of relevant domain, (b) the choice of irrelevant domain, and (c) the choice of time delay between training program and transfer measure administration.

Of the six courses within the managerial training program, Marketplace Awareness was designated the relevant domain and evaluated in this study. However, this course was not selected because it was the optimal course to evaluate, but because delays in the implementation of subsequent courses prevented their selection. In fact, the declarative knowledge-based content of Marketplace Awareness actually hindered implementation of the IRS approach in the transfer evaluation. This is best illustrated by comparing the item writing process for knowledge and behavioral items. With Marketplace Awareness, the process of writing relevant knowledge items was relatively easy since

they could be extracted from the factual knowledge present in the course training material. In a similar manner, irrelevant knowledge items were extracted from the factual knowledge present in Data Analysis, a subsequent course (i.e., the irrelevant domain). However, because Marketplace Awareness did not provide any direct skills training, the process of writing relevant behavioral items for the transfer measure proved more difficult. Ultimately, relevant behavioral items were written to reflect the ways in which managers might apply the knowledge they had learned. For example, having learned that the telecommunications industry is highly competitive, managers might apply this knowledge by developing and employing customer service and retention skills. The process of writing irrelevant behavioral items proved to be easier, as they were simply drawn from two subsequent training courses (Data Analysis and Front-line Leadership), both of which provided skills-oriented training.

As this study demonstrated, use of the IRS approach to evaluate a declarative knowledge-based training program in terms of trainee transfer of training is not impossible, only more difficult than a training program providing instruction in both knowledge and skills. This problem was not wholly unexpected, as Haccoun and Hamtiaux (1994) hinted that their learning investigation provided no information as to the suitability of an

IRS approach for evaluating training programs in terms of trainee behaviors. Although they believed the approach could be used with behavioral measurements, they warned "the test construction problem may be more complex" (p. 603).

Time constraints also prompted the study's choice of irrelevant domain. When choosing an irrelevant domain from which to construct irrelevant items, Haccoun and Hamtiaux (1994) suggested that irrelevant items be drawn from training objectives or training content that logically could have been included in the training program being evaluated, but for whatever reason, were not. Taken literally, this would require that the training evaluator communicate with the instructional designers responsible for creating and editing the course, identify training content which was excluded from the course, and extract irrelevant items from this content for use in the evaluation.

Unfortunately, time constraints required the rapid construction of items and restricted this communication with instructional designers. Furthermore, the instructional designers were not completely supportive of efforts to evaluate their course. Consequently, the irrelevant domain consisted of training content derived from two subsequent managerial training courses (i.e., Data Analysis and Front-line Leadership), both of

which were linked to the relevant domain (i.e., the Marketplace Awareness) by the results of a single training needs assessment. As previously stated, this method may have contributed to a Type II error in the presence of a maturation-history interaction, and is therefore a less desirable alternative to Haccoun and Hamtiaux's (1994) technique.

Finally, practical constraints also directed the length of the delay between the training program and the administration of the transfer measure. When evaluating transfer of training, it is necessary to include a delay of an appropriate length in order to allow time for trainees to develop and demonstrate what they had learned in training on the job. The length of this delay is a critical aspect of any transfer evaluation, as without an adequate delay, the transfer evaluation may demonstrate little or no transfer. Tannenbaum and Yukl (1992) have stated that there is no single formula or set of guidelines for determining an appropriate delay. However, in general, the length of the delay depends on the complexity of the behaviors to be transferred. For example, Cascio (1991) suggested a delay of at least 3 months when evaluating the transfer of complex behaviors (e.g., decision-making skills).

In practice, the length of delay is determined after the target behaviors are identified. Unfortunately, in this study, organizational

constraints demanded a 2 week delay, as the evaluation had to occur prior to the delivery of a subsequent course. Therefore, the transfer measure could only include those behaviors which trainees could reasonably be expected to demonstrate in that period of time. This constraint ultimately affected both the quantity and quality of the behavioral items included in the transfer measure.

Future Research

Directions for future research include (a) modifying this research design based on lessons learned, (b) investigating alternative methods of generating relevant and irrelevant items, and (c) investigating alternative sources of transfer ratings.

Given the impact of organizational constraints on the implementation of this study, future researchers should consider modifying this research design based on several lessons learned. First, researchers should select an appropriate relevant and irrelevant domain for evaluation. Not only will this ease the process of behavioral item writing in a transfer of training evaluation, but it may also lesson the risk of a maturation-history interaction. Second, researchers should conduct a pilot test of relevant and irrelevant items prior to the evaluation. By

removing those items exhibiting ceiling effects, the IRS approach will be less vulnerable to a Type II error. Third, allow the target behaviors to determine the appropriate delay between training program and the administration of the transfer measure. If the "tail is allowed to wag the dog" (i.e., a given delay is used to select appropriate behaviors), the choice of target behaviors may lack both quantity and quantity. Finally, embed the IRS approach within a more traditional two-group pretest-posttest design. By using a more rigorous design employing a control group, the analysis is likely to produce more comprehensive results and more confident conclusions.

Future research might also investigate alternative methods of generating relevant and irrelevant items. For example, rather than extracting irrelevant items--each assumed to be equally irrelevant--from an irrelevant domain (i.e., material that logically could have been included in the training), the researcher could create a variety of items exhibiting several degrees of relevancy (or irrelevancy). Instructional designers or subject matter experts could rate the relevancy (or irrelevancy) of each item and the training evaluator could then use these items in an IRS design.

The source of transfer ratings might also be worthy of exploration. In most organizational settings, it is typically the supervisor's responsibility to provide judgments of the adequacy of a subordinate's performance for the purposes of performance appraisal (Bernardin & Beatty, 1984; Campbell et al., 1970; Kraiger, 1985; Klimoski & London, 1974; Lacho, Stearns, & Villere, 1979; Malka, 1990; Mount, 1984) and training needs assessment (Campbell et al., 1970; Cleveland, Murphy, & Williams, 1989; Levine, 1986; McEnery & McEnery, 1987; Yammarino & Atwater, 1993). However, in recent years, the incumbent has increasingly been recommended as a data source (e.g., Hazucha, Hezlett, & Schneider, 1993; London & Beatty, 1993; Van Velsor, Taylor, & Leslie, 1993; Yammarino & Atwater, 1993). Choice of rating source could be a critical issue given that individuals often have a significantly different view of their own job performance as compared to those of their supervisors (Baird, 1977; Heneman, 1974; Thornton, 1968, 1980; Staley & Shockley-Zalabak, 1986).

It might be interesting to learn how each source's characteristic biases affect the results of a transfer evaluation involving behavioral ratings. Self-ratings are prone to leniency; reviews and meta-analyses of literature concerning self-ratings in performance appraisal have concluded

that, on average, individuals rate themselves higher than other rating groups (Harris & Schaubroeck, 1988; Kraiger, 1985; Thornton, 1980). Supervisors' ratings manifest more halo error than self-ratings (Thornton, 1980; Kraiger, 1985), as supervisors evaluate according to an overall impression rather than by distinguishing among individual items (Vance, Winnie, & Wright, 1983; McEnery & McEnery, 1987; Malka, 1990).

It is perhaps fortunate that self-ratings rather than supervisory ratings where available for this study, as there is some evidence that self-ratings, although lenient, may still be useful for evaluating the transfer of training. For example, Mabe and West (1982) suggested ". . . self-enhancing, accurate, or modest reports may be found, depending on certain conditions" (p. 287). In their review and meta-analysis of self-evaluation studies, Mabe and West found that leniency errors were reduced under certain conditions, such as when measures were anonymous, specific, and separated in time from performance measures used for rewards and promotion, and when participants were intelligent, experienced, and believed that there was little to be gained from self-enhancement. Thornton (1980) provided some evidence for the latter condition by concluding that there is less leniency in self-ratings obtained under "research only" conditions. Others (e.g., Arnold & Davey, 1992;

McEnery & McEnery, 1987; Somers & Birnbaum, 1991) have since echoed Mabe and West's claims. Thus, self-ratings may have been appropriate for this study, as the transfer measure was an anonymous, behaviorally specific checklist, used only for the purposes of training evaluation. In future research, it would be useful to determine the comparative capability of self- and supervisory measures of behavior to assess transfer of training.

Conclusions

This study is the first attempt at applying the IRS approach to the evaluation of transfer of training in a field setting. Despite the approach's initial failure to detect trainee transfer, the subsequent analysis demonstrated its utility after threats to internal validity had been properly controlled. Therefore, training researchers and practitioners should consider the IRS approach when seeking a rigorous, yet easily implemented, method for conducting training evaluations in field settings.

REFERENCES

Adams, J. (1987). Historical review and appraisal of research on learning, retention, and transfer of human motor skills. Psychological Bulletin, 101(1), 41-74.

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. Personnel Psychology, 42(2), 331-342.

Alliger, G. M., Tannenbaum, S. I., & Bennett, W. (1995, August). A meta-analysis of relations among training criteria. Paper presented at the annual meeting of the American Psychological Association, New York.

Arnold, J., & Davey, K. M. (1992). Self-ratings and supervisor ratings of graduate employees' competencies during early career. Journal of Occupational and Organizational Psychology, 65(3), 235-250.

Arvey, R. D., & Cole, D. A. (1989). Evaluating change due to training. In I. L. Goldstein (Ed.), Training and development in organizations. San Francisco: Jossey-Bass.

Baird, L. S. (1977). Self- and supervisor rating of performance: As related to self-esteem and satisfaction with supervision. Academy of Management Journal, 20(2), 291-300.

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. Personnel Psychology, 41(1), 63-105.

Baldwin, T. T., Magjuka, R. G., & Loher, B. T. (1991). The perils of participation: Effects of choice of training on trainee motivation and learning. Personnel Psychology, 44(1), 51-65.

Beaudin, B, P. (1987). Enhancing the transfer of job-related learning from the learning environment to the workplace. Performance and Instruction, 26(9-10), 19-21.

Bernardin, H. J., & Beatty, R. W. (1984). Performance appraisal: Assessing human behavior at work. Boston: Kent.

Blaiwes, A. S., Puig, J. A., & Regan, J. J. (1973). Transfer of training and the measurement of training effectiveness. Human Factors, 15(6), 523-533.

Bunker, K. A., & Cohen, S. L. (1977). The rigors of training evaluation: A discussion and field demonstration. Personnel Psychology, 30(4), 525-541.

Camp, R. R., Blanchard, P. N., & Huszczo, G. E. (1986). Toward a more organizationally effective training strategy and practice. Englewood Cliffs, NJ: Prentice-Hall.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., III, & Weick, K. E., Jr. (1970). Managerial behavior, performance and effectiveness. New York: McGraw-Hill.

Cannon-Bowers, J. A., Salas, E., Tannenbaum, S. I., & Mathieu, J. E. (1995). Toward theoretically based principles of training effectiveness: A model and initial empirical investigation. Military Psychology, 7(3), 141-164.

Cascio, W. F. (1991). Applied psychology in personnel management (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Cascio, W. F., & Zammuto, R. F. (1989). Societal trends and staffing policies. In W. F. Cascio (Ed.). Human resource planning, employment, and placement (pp. 2-1 to 2-33). Washington DC: Bureau of National Affairs.

Clark, C. S., Dobbins, G. H., & Ladd, R. T. (1993). Exploratory field study of training motivation. Group and Organizational Management, 18(3), 292-307.

Clement, R. W. (1982). Testing the hierarchy theory of training evaluation: An expanded role for trainee reactions. Public Personnel Management Journal, 11(2), 176-184.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. Journal of Applied Psychology, 74(1), 130-135.

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

Cook, T. D., Campbell, D. T., & Perracchio, L. (1990). Quasi experimentation. In M. D. Dunnette, & L. M. Hough (Eds.), Handbook of industrial and organizational psychology: Vol. 1. (2nd ed., pp. 491-576). Palo Alto, CA: Consulting Psychologists Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometricka, 16, 297-334.

Davis, R. H. (1979). A behavioral change model with implications for faculty development. Higher Education, 8, 123-140.

Facteau, J. D., Dobbins, G. H., Russell, J. E. A., Ladd, R. T., & Kudisch, J. D. (1995). The influence of general perceptions of the training environment on pretraining motivation and perceived training transfer. Journal of Management, 21(1), 1-25.

Ferguson, W. C. (1968). Quantitative evaluation of training using student reaction. Training and Development Journal, 22(11), 36-43.

Finkel, C. L. (1987). The true cost of a training program. Training and Development Journal, 41(9), 74-76.

Ford, J. K., & Kraiger, K. (1995). The applications of cognitive psychology constructs and principles to the instructional systems model of training: Implications for needs assessment, design, and transfer. International Review of Industrial and Organizational Psychology, 10, 1-48.

Ford, J. K., Quinones, M. A., Sego, D. J., & Sorra, J. S. (1992). Factors affecting the opportunity to perform trained tasks on the job. Personnel Psychology, 45(3), 511-527.

Fullerton, H. N., Jr. (1989). New labor force projections, spanning 1988 to 2000. Monthly Labor Review, 112(11), 3-12.

Fullerton, H. N., Jr. (1995). The 2005 labor force: Growing, but slowly. Monthly Labor Review, 118(11), 29-44.

Geber, B. (1995). Does your training make a difference? Prove it! Training, 32(3), 27-34.

Georgenson, D. L. (1982). The problem of transfer calls for partnership. Training and Development Journal, 36(10), 75-78.

Gist, M. E., Bavetta, A. G., & Stevens, C. K. (1990). Transfer training method: Its influence on skill generalization, skill repetition, and performance level. Personnel Psychology, 43(3), 501-523.

Gist, M. E., Stevens, C. K., & Bavetta, A. G. (1991). Effects of self-efficacy and post-training intervention on the acquisition and maintenance of complex interpersonal skills. Personnel Psychology, 44(4), 837-861.

Goldstein, I. L. (1980). Training in work organizations. Annual Review of Psychology, 31, 229-272.

Goldstein, I. L. (1993). Training in organizations: Needs assessment, development, and evaluation (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Goldstein, I. L., & Gilliam, P. (1990). Training system issues in the year 2000. American Psychologist, 45(2), 134-143.

Gorman, C. (1988, December 19). The literacy gap. Time, 132(25), 56-57.

Grove, D. A., & Ostroff, C. (1990). Program evaluation. In K. N. Wexley & J. Hinrichs (Eds.), Developing human resources. Washington, DC: BNA Books.

Haccoun, R. R., & Hamtiaux, T. (1994). Optimizing knowledge tests for inferring learning acquisition levels in single group training effectiveness designs: The internal referencing strategy. Personnel Psychology, 47(3), 593-604.

Hamblin, A. C. (1974). Evaluation and control of training. New York: McGraw-Hill.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personnel Psychology, 41(1), 43-62.

Hazucha, J. F., Hezlett, S. A., & Schneider, R. J. (1993). The impact of 360-degree feedback on management skills development. Human Resource Management, 32(2 & 3), 325-351.

Heneman, H. G., III. (1974). Comparison of self- and supervisor ratings of managerial performance. Journal of Applied Psychology, 59(5), 638-642.

Hicks, W. D., & Klimoski, R. J. (1987). Entry into training programs and its effects on training outcomes: A field experiment. Academy of Management Journal, 30(3), 542-552.

Huczynski, A. A., & Lewis, J. W. (1980). An empirical study into the learning transfer process in management training. Journal of Management Studies, 17(2), 227-240.

Kelly, H. B. (1982). A primer on transfer of training. Training and Development Journal, 36(11), 102-106.

Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig (Ed.), Training and development handbook (2nd ed., pp. 18-1 to 18-27). New York: McGraw-Hill.

Kirkpatrick, D. L. (1978). Evaluating in-house training programs. Training and Development Journal, 32(9), 6-9.

Klein, K. J., & Hall, R. J. (1988). Innovations in human resource management: Strategies for the future. In J. Hage (Ed.), Future of organizations. (pp. 147-162). Lexington, MA: Lexington Books.

Klimoski, R. J., & London, M. (1974). Role of the rater in performance appraisal. Journal of Applied Psychology, 59(4), 445-451.

Kraiger, K. (1985, September). Analysis of relationships among self-, peer, and supervisory ratings of performance. Denver: University of Colorado, Department of Psychology.

Kraiger, K. (1995, August). Implications of cognitive psychology for training research: A tail of two paradigms. Invited address at the annual meeting of the American Psychological Association, New York.

Lacho, K. J., Stearns, G. K., & Villere, M. R. (1979). A study of employee appraisal systems of major cities in the United States. Public Personnel Management, 8(2), 111-125.

Leifer, M. S., & Newstrom, J. W. (1980). Solving the transfer of training problems. Training and Development Journal, 34(9), 42-46.

Levine, H. Z. (1986). Performance appraisals at work. Personnel, 63(6), 63-71.

London, M., & Beatty, R. W. (1993). 360-degree feedback as a competitive advantage. Human Resource Management, 32(2 & 3), 353-372.

McEnery, J., McEnery, J. M. (1987). Self-rating in management training needs assessment: A neglected opportunity? Journal of Occupational Psychology, 60(1), 49-60.

Mabe, P. A., III, & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. Journal of Applied Psychology, 67(3), 280-296.

Malka, S. (1990). Application of the multitrait-multirater approach to performance appraisal in social service organizations. Evaluation and Program Planning, 13(3), 243-250.

Mathieu, J. E., Tannenbaum, S. I., & Salas, E. (1992). Influences of individual and situational characteristics on measures of training effectiveness. Academy of Management Journal, 35(4), 828-847.

Mosel, J. N. (1957). Why training programs fail to carry over. Personnel, 34(3), 56-64.

Mount, M. K. (1984). Psychometric properties of subordinate ratings of managerial performance. Personnel Psychology, 37(4), 687-702

Newstrom, J. W. (1978). Catch-22: The problems of incomplete evaluation of training. Training and Development Journal, 32(11), 22-24.

Noe, R. A. (1986). Trainees' attributes and attitudes: Neglected influences on training effectiveness. Academy of Management Review, 11(4), 736-749.

Noe, R. A., & Ford, J. K. (1992). Emerging issues and new directions for training research. In G. Ferris & K. Rowland (Eds.), Research in personnel and human resources management (Vol. 10, pp. 345-384). Greenwich, CT: JAI Press.

Noe, R. A., & Schmitt, N. (1986). The influence of trainee attitudes on training effectiveness: Test of a model. Personnel Psychology, 39(3), 497-521.

Ostroff, C. (1991). Training effectiveness measures and scoring schemes: A comparison. Personnel Psychology, 44(2), 353-374.

Quinones, M. A. (1995). Pretraining context effects: Training assignment as feedback. Journal of Applied Psychology, 80(2), 226-238.

Ralphs, L. T., & Stephan, E. (1986). HRD in the Fortune 500. Training and Development Journal, 40(10), 69-76.

Russell, J. S., & Wexley, K. N. (1988). Improving managerial performance in assessing needs and transferring training. Research in Personnel and Human Resources Management, 6, 289-324.

Saari, L. M., Johnson, T. R., McLaughlin, S. D., & Zimmerle, D. M. (1988). A survey of management training and education practices in U.S. companies. Personnel Psychology, 41(4), 731-743.

Sackett, P. R., & Mullen, E. J. (1993). Beyond formal experimental design: Towards an expanded view of the training process. Personnel Psychology, 46(3), 613-627.

Salas, E., Burgess, K. A., & Cannon-Bowers, J. A. (1995). Training effectiveness techniques. In J. Wiener (Ed.), Research techniques in human engineering (pp. 439-471). Englewood, NJ: Princeton & Hall.

Smith, E. M., Ford, J. K., Weissbein, D. A., & Gully, S. M. (1995, May). The effects of goal orientation, metacognition, and practice strategies on learning and transfer. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Somers, M. J., & Birnbaum, D. (1991). Assessing self-appraisal of job performance as an evaluation device: Are the poor results a function of method or methodology? Human Relations, 44(10), 1081-1091.

Staley, C. C., & Shockley-Zalabak, P. (1986). Communication proficiency and future training needs of the female professional: Self-assessment vs. supervisors' evaluations. Human Relations, 39(10), 891-902.

Swierczek, F. W., & Carmichael, L. (1985). The quantity and quality of evaluation training. Training and Development Journal, 39(1), 95-99.

Tannenbaum, S. I., Mathieu, J. E., Salas, E., & Cannon-Bowers, J. A. (1991). Meeting trainees' expectation: The influence of training fulfillment on the development of commitment, self-efficacy, and motivation. Journal of Applied Psychology, 76(6), 759-769.

Tannenbaum, S. I., & Woods, S. B. (1992). Determining a strategy for evaluating training: Operating within organizational constraints. Human Resource Planning, 15(2), 63-81.

Tannenbaum, S. I., & Yukl, G. (1992). Training and development in work organizations. Annual Review of Psychology, 43, 399-441.

Thayer, P. W. (1997). A rapidly changing world: Some implications for training systems in the year 2001 and beyond. In M. A. Quinones & A. Ehrenstein (Eds.), Training for a rapidly changing workplace (pp. 15-30). Washington, DC: American Psychological Association.

Thornton, G. C., III. (1968). Relationship between supervisory and self-appraisals of executive performance. Personnel Psychology, 21(4), 441-455.

Thornton, G. C., III. (1980). Psychometric properties of self-appraisals of job performance. Personnel Psychology, 33(2), 263-271.

Tracey, J. B., Tannenbaum, S. I., & Kavanagh, M. J. (1995). Applying trained skills on the job: The importance of the work environment. Journal of Applied Psychology, 80(2), 239-252.

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

Trulson, M. E. (1986). Martial arts training: A novel "cure" for juvenile delinquency. Human Relations, 39(12), 1131-1140.

Turnage, J. J. (1990). The challenge of new workplace technology for psychology. American Psychologist, 45, 171-178.

Tziner, A., Haccoun, R. R., & Kadish, A. (1991). Personal and situational characteristics influencing the effectiveness of transfer of training improvement strategies. Journal of Occupational Psychology, 64(2), 167-177.

Vance, R. J., Winnie, P. S., & Wright, E. S. (1983). A longitudinal examination of rater and ratee effects in performance ratings. Personnel Psychology, 36(3), 609-620.

Van Velsor, R., Taylor, S., & Leslie, J. B. (1993). An examination of the relationships among self-perception accuracy, self-awareness, gender, and leaders effectiveness. Human Resource Management, 32(2 & 3), 249-263.

Weiss, H. M. (1990). Learning theory and industrial psychology. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology: Vol. 1. (2nd ed., pp. 171-221). Palo Alto, CA: Consulting Psychologists Press.

Wexley, K. N., & Baldwin, T. T. (1986). Posttraining strategies for facilitating positive transfer: An empirical exploration. Academy of Management Journal, 29(3), 503-520.

Wexley, K. N., & Latham, G. P. (1991). Developing and training human resources in organizations (2nd ed.). New York: Harper-Collings.

Yammarino, F. J., & Atwater, L. E. (1993). Understanding self-perception accuracy: Implications for human resource management. Human Resource Management, 32(2 & 3), 231-247.

Yelon, S. (1992). M.A.S.S.: A model for producing transfer. Performance Improvement Quarterly, 5(2), 13-23.