Order Number 9130319

The effectiveness of performance appraisal training: Alpha, beta, and gamma congruence

Gracin, Lynn, Ph.D.

City University of New York, 1991

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

THE EFFECTIVENESS OF PERFORMANCE APPRAISAL TRAINING: ALPHA , BETA, AND GAMMA CONGRUENCE

by

LYNN GRACIN

A dissertation submitted to the Graduate Faculty in Psychology
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy, The City University of New York.

1991

This manuscript has been read and accepted by the Graduate Faculty in Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

4/19/91
Date

Roger E. Millsap
Chair of Examining Committee

4/19/91
Date

Herbert D. Saltzstein
Executive Officer

Joel Lefkowitz

Walter Reichman

Ramon M. Henson

David A. Rodriguez

Supervisory Committee

The City University of New York

Abstract

THE EFFECTIVENESS OF PERFORMANCE APPRAISAL TRAINING: ALPHA, BETA, AND GAMMA CONGRUENCE

by

Lynn Gracin

Adviser: Professor Roger E. Millsap

The purpose of this study was to introduce the proposed constructs of beta and gamma congruence and to assess promising performance appraisal paradigms, Frame of Reference (FOR) and Rater Error Training (RET), in light of these proposed constructs. Historically, these training paradigms have been assessed at an alpha level, that is, the correspondence of observed performance ratings with true score estimates provided by expert raters. Alpha congruence however presumes a common expert--trainee metric (beta congruence) and conceptual domain (gamma congruence). Expert ratings were obtained from nine (9) Ph.D. I/O Psychologists and sixty-five (65) I/O Psychology doctoral students. Twelve undergraduate classes (236 students) were randomly assigned to one of three conditions: Frame of

Reference Training, Frame of Reference and Rater Error Training (FRET), and no training (Comparison).

Gamma congruence was found for both the FOR and comparison groups. The FRET group did not exhibit gamma congruence. These findings demonstrate that FOR training may not result in more accurate frame utilization than is evidenced by untrained raters, and that RET training may decrease accuracy in conceptual domains.

Beta congruence was absent for the FOR and comparison group. High correlations among the factors signaled that trainees perceived less differentiation among the relevant constructs than experts. The effect of RET on scale calibration could not be examined since members of the FRET group were not evaluating the same constructs as the expert group.

Traditional accuracy assessments of elevation and distance accuracy, analogous to alpha level assessments, revealed that the trained groups (FOR, FRET) were significantly different from the comparison group. These findings are explained with regard to relative rather than absolute accuracy. It is concluded that traditional alpha congruence assessments are inappropriate and misleading in the absence of beta and gamma congruence.

Acknowledgements

Upon completion of a dissertation one recognizes the many individuals in their life which assisted in the process. I would like to thank those people...

First and foremost I would like to thank my mother, Ada Gracin whose constant love, support, and belief in me empowered me to undertake a doctorate degree. Mom, Harding Cotter said that "There are only two lasting bequests we can hope to give our children. One of these is roots; the other, wings". Thanks for both and more, I love you!

Many thanks to my other family members, Hank, Jay, Grandma Minnie, Marisol, and Julie for their understanding and flexiblility in accomodating my maniacal time commitments that were channeled toward completing this degree.

I would also like to acknowledge my dissertation committee members:

Roger E. Millsap your wisdom, energy, guidance and friendship not only eased much of the pain of the dissertation process but made this posssible. Thanks for always being there.

Joel Lefkowitz, thanks for accepting me in to the program and teaching me my most valuable graduate school experience lesson. You taught me the value of critical thinking and gave me the confidence to exercise those thoughts. Thanks.

Water Reichman, your emotional support, professional guidance, and belief in me brought much happiness to my graduate school days, thanks.

Ray Henson, thank you for taking the time from your very busy schedule to participate in my defense. Your professionalism and expertise is extremely valued.

David Rodriguez, thanks for finally coming to the defense...please stay healthy and beware of elevators. Honestly, I want to thank you for never letting me forget that my dissertation was top priority even though it may have been in conflict with your past role as my supervisor. Your support and friendship is treasured.

I would additionally like to thank Tammi Brady who took much time from her busy schedule to offer both her technical and emotional support.

Many thanks to my friends and colleagues at Baruch; Sandy Hartog, Lori Poveromo, and Margaret Geoghan who always had an ear for me and provided great support and many good times.

Gautier said "To love is to admire with the heart; to admire is to love with the mind". Last, but not least, I would like to thank the man I both love and admire, Dr. Nicholas C. Stilwell. Nick, thanks for so patiently standing behind me and encouraging me to exercise my potential, I love you!

Table of Contents

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

List of Tables

List of Tables (continued)

CHAPTER I

The following quote from Robert Wherry cited in Bernardin (1984) best describes the short history of performance appraisal training research: "We don't know what we're doing but we're doing it very carefully and hope you are pleased with our unintelligent diligence".

INTRODUCTION

The utilization of performance ratings is extremely widespread. Many organizations rely on performance evaluations at least annually for a variety of administrative and developmental decisions: transfers, allocation of financial rewards, developmental and training needs. Their frequency is not limited to organizational life, they are also utilized extensively for research purposes. In fact, in the mid-sixties, Guion (1965) found that over three-fourths (81%) of the studies published in the Journal of Applied Psychology and Personnel Psychology used some form of rating as a criterion. Fifteen years later, Landy and Trumbo (1980) found that in a decade of Journal of Applied Psychology articles (1965-1975), ratings were used as criteria in seventy-two percent of the validation studies. Although the above citations include both ratings of objective standards (i.e., number of widgets produced, sales performance) and judgmental performance ratings (i.e., communication skills), there is no doubt that the use of judgmental performance ratings is extremely prevalent. Because of their popularity as criteria and tools for organizational decisions, a

great deal of research has been generated due to concerns with these measures.

Evaluation Indices

The majority of reservations associated with the use of performance appraisals reflect concerns with rater subjectivity. Appraisal judgments have been examined by evaluating rater error (halo, leniency/severity, central tendency and restriction of range) and/or accuracy. As described below, many researchers have criticized and abandoned error evaluation and instead have recently focused on accuracy. Accuracy however, as currently defined and applied is also problematic.

Halo, leniency/severity, central tendency and restriction of range are the most commonly mentioned errors in the performance appraisal literature. The value of these indexes have been questioned theoretically and methodologically by Sulsky & Balzer (1988). Theoretical reservations center around the possibility that halo and other errors in part reflect reality. True trait inter-correlation is possible, if not likely. For example, an individual who has good

leadership skills may be likely to possess good interpersonal skills, oral skills, and effective management processes (planning, organizing, controlling). If a performance appraisal scale consisted of the above dimensions and the ratee exhibited high leadership ability, the ratings would be relatively invariant and positively skewed. It might then be concluded that the ratings possess halo, leniency, and restriction of range. Thus, vis-a-vis error assessment, the evaluation would not be considered indicative of performance, when in fact the ratings reflect reality. Even further complicating the evaluation of training using error as criteria is the methodology used for evaluating error. There are multiple operational definitions employed for each of the error measures (Bernardin & Beatty, 1984; Saal, Downey, & Lahey, 1980), and in fact, the varied definitions have been found to lead to different conclusions (Murphy & Balzer, 1981; Saal, Downey, & Lahey, 1980). Fraught with these problems, many researchers have abandoned the practice of evaluating error and instead have focused on the accuracy of ratings. However, accuracy as defined below, is not

the sole "crucial criterion in judging performance rating quality (Pulakos, 1986, p.77)".

Theoretically, the construct of accuracy has great value for evaluating the impact of performance appraisal training; the more accurate our ratings, theoretically the more effective our decision making and valuable our research (when ratings are used as criteria). Although accuracy has implicitly been a goal of performance appraisal research, during the period when only error was assessed, it was just assumed that less error meant greater accuracy. This relationship is however not found with the current operationalization of error and accuracy (Bernardin & Pence, 1980). Error and accuracy are currently measured as two different concepts with two different operationalizations. Error is based on an assumed distribution and accuracy has been defined empirically (i.e., expert ratings). Error and accuracy can only meaningfully be compared in a situation in which they are based on the same distribution of scores (e.g., expert true scores). Furthermore, methodologically, not unlike error, there are multiple methods of evaluating accuracy. The commonality among these

methods is the assessment of absolute level of correspondence between true score estimates provided by "expert" raters and other raters' ratings. The closer the correspondence, the more accurate the ratings are thought to be (Borman, 1977). Specifically, there are six types of accuracy discussed in the performance appraisal literature: D^2, elevation, differential elevation, stereotype accuracy, differential accuracy, distance accuracy and correlation accuracy.

Prior to 1955, accuracy in the judgment literature was assessed as D^2, the sum of the squared differences between subject ratings and true score estimates across ratees and dimensions. Cronbach's (1955) seminal article highlighted the fact that D^2 ignored important accuracy information and should be partitioned into separate components to facilitate interpretation. Specifically, D^2, according to Cronbach (1955) is comprised of the following four components: a) elevation (E), b) differential elevation (DE), c) stereotype accuracy, d) differential accuracy (Cronbach, 1955). These components of D^2, described below, are the criteria referenced and utilized for assessment of performance appraisal rating accuracy.

Elevation accuracy is the extent to which a rater approximates true score estimates of the overall performance of ratees (Roach & Gupta, 1990). Specifically, elevation accuracy is the difference between a rater's overall mean rating and the true overall mean (as established by expert raters). Differential elevation, stereotype accuracy, and differential accuracy, are usually expressed in analysis of variance (ANOVA) terms (Sulsky & Balzer, 1988; Fisicaro, 1988). Differential elevation is the extent to which a rater approximates true score estimates for each ratee, collapsed across dimensions (Roach & Gupta, 1990). In an ANOVA it is the differential main effect of ratees. Stereotype accuracy is the extent to which a rater approximates true score estimates for each performance dimension across ratees, in other words, it is the main effect of dimensions. Differential accuracy is the extent to which raters approximate true score estimates of differences, averaged across dimensions, between ratees on dimensions (Roach & Gupta, 1990). Thus, in ANOVA terms it is a ratee by dimension interaction.

Borman (1977) was the first to apply the phenomenon

of accuracy to the performance appraisal literature. He felt that differential accuracy was the most appropriate for assessing performance judgments since it provided information on the rank ordering of target persons by dimension. Accurately discriminating among ratees on a number of performance dimensions is certainly meaningful but is it the most meaningful? Proper rank ordering of ratees is apparently important for between-ratee judgments such as promotions and salary increases; but within ratee judgments of relative strengths, weaknesses and developmental needs are equally important in the appraisal context. Differential accuracy assessments do not address the accuracy of these important within ratee judgments. Nevertheless, Borman's DA measure is conceptually similar yet statistically different than Cronbach's. Instead of considering the distance between raters' and experts' true score estimates, Borman's measure looks at the correlation between the two. McIntyre (1984) has termed Borman's DA, correlational accuracy, since it is the average correlation of rater and true score estimates by dimension across ratees. McIntyre, Smith, & Hassett (1984) and Athey & McIntyre (1987) adopted a

modified form of Cronbach's D^2 index, termed distance accuracy (DA). Distance accuracy is based on Davidson' (1985) city block metric and reflects the average absolute deviation of subject ratings from true scores. Much of the judgment literature employs one or more of the above accuracy evaluations as criteria for evaluation. All accuracy assessments employ as the operationalization of accuracy one or more measures of observed score rater and expert correspondence, specifically D^2, elevation, differential elevation, stereotype accuracy, differential accuracy, distance accuracy, and/or correlation accuracy.

Performance Appraisal Training

Rater Error Training

Although research on performance appraisal training is fairly substantial, little can be concluded about its usefulness. The training literature has, by-and-large, consisted of error or accuracy evaluations of three types of training: rater error training (RET), rater accuracy training (RAT), and frame of reference training (FOR). Early training focused on diminishing rater errors in order to enhance ratings. Rater error

training simply involves the presentation and or discussion of definitions, illustrations, or examples of rater errors (i.e., halo, leniency, central tendency) to trainees and instructions not to commit them. It attempts to train raters to change their rating distributions. For example, trainees are instructed to rate all dimensions separately, not to give the same ratings for all or most items, to avoid giving all ratings that fall in the middle of the scale, and to avoid giving ratings that are more favorable or severe than those deserved. Although the core of RET training has the above content, the methods of training are diverse and varied. Researchers have presented definitions, illustrations and graphic examples of rater errors (Bernardin & Walter, 1977; Bernardin, 1978; Ivancevich, 1979; McIntyre et al., 1984), others have also conducted group discussions (Bernardin, 1978; Bernardin & Pence, 1980, Latham, Wexley, & Pursell, 1975) and some have provided opportunities for practice and feedback (Ivancevich, 1979; Spool, 1979). Despite the process differences, all provide participants with information about rater error and the implication is that certain rating

distributions are more desirable than others. In general, the more active the raters are in the training process, the greater the outcome (i.e., decreased rater error) (Smith,1986). Specifically, the opportunity to participate in practice and feedback produces better results than presenting a lecture (Smith, 1986).

For many years, it was assumed that the more accurate ratings are those that have less psychometric error. RET has been found to reduce rater error (Bernardin, 1978; Ivancevich, 1979; Warmke & Billings, 1979) and to result in higher scores on knowledge tests of psychometric error (Bernardin, 1978). However, RET training has not lead to greater differential, distance, or correlational accuracy (Borman, 1975; Borman, 1979; Bernardin & Pence, 1980; Pulakos, 1984; McIntyre, et al., 1984). In fact, it has been suggested that RET fosters a response set that decreases accuracy (Bernardin & Pence, 1980); in an effort to spread out ratings, inaccuracies may be introduced. Why would information about rating errors lead to greater observed score correspondence between trainees and experts (i.e., accuracy as currently defined in the performance appraisal literature)?

Accuracy, unlike error, is defined by an independent set of expert ratings. The intent of RET is for trained raters to properly utilize the presented rating scale, for instance, not to be too lenient, severe, or restrict the range of ratings. The utility of RET for improving ratings can more appropriately be assessed by examining whether trained raters appropriately use the rating scales. Clearly, current accuracy assessments are not appropriate for measuring RET training utility.

Rater Accuracy and Frame of Reference Training

The RET evaluation results generated new and different training emphases. Researchers began to investigate rater accuracy training (RAT) and frame of reference (FOR) training. RAT training is similar to RET in that the multidimensionality of jobs and the need to distinguish between dimensions is emphasized. Unlike RET, no assumed distribution is presented and there is a greater focus on ratee behavior. Specifically, rater accuracy training typically consists of a lecture on the multidimensionality of jobs, the need to distinguish between dimensions, and the need to pay close attention to performance in light

of these dimensions (Bernardin & Pence, 1980; Pulakos, 1984). Pulakos's (1984) RAT training is different from Bernardin & Pence's (1980) in that it attempts to provide a frame of reference by giving participants anchors of expected performance at varied scale levels. Pulakos's (1984) work is not based on traditional RAT nor current FOR training, it is a hybrid of the two. FOR training evolved from the RAT work. Like RAT, FOR focuses on learning correct performance standards to facilitate agreement in evaluating behaviors. FOR however, unlike RAT, does not require the trainees to attain a frame of reference through their own efforts. Instead the frame of reference or schema is established by expert raters and presented to the trainees (McIntyre et al., 1984; Athey & McIntyre, 1987). FOR training suggests that raters with more valid schemata, cognitive organizing structures, will be more sensitive to relevant ratee behaviors and thus rate more accurately. FOR training as proposed by Bernardin and Buckley (1981), the individuals responsible for its introduction (although inspired by Borman (1979)), consists of a job description for trainees, a discussion of duties and qualifications, vignettes of

critical incidents of job performance, trainee ratings and justification of each vignette, trainer feedback of "correct" ratings, and a discussion focusing on discrepancies. FOR as practiced to date (McIntyre, Smith, & Hassett, 1984; Athey & McIntyre, 1987) does not include all of the above, instead the core focus is on providing trainees with expert schemata to sensitize them to relevant ratee behaviors. Given the current "cognitive revolution" (Landy, 1985), this new emphasis was appealing to many. However, the evaluations of RAT and FOR training do not lead to compelling positive conclusions.

RAT training has been found to result in lower deviations from dimension "true" rating estimates than RET groups (Bernardin & Pence, 1980) and greater correlational accuracy (Pulakos, 1985). Greater accuracy than RET groups is not surprising especially in light of the above discussion. In addition, RAT groups have not outperformed control groups with respect to accuracy (Bernardin & Pence,1980).

FOR training as examined by McIntyre and his colleagues (McIntyre, Smith, & Hassett, 1984; Athey & McIntyre, 1987) has not included all the components as

suggested by Bernardin and Buckley (1981). McIntyre and his colleagues (McIntyre, Smith, & Hassett, 1984; Athey & McIntyre, 1987) provided the three components that they felt were basic to FOR training: information describing the job, practice and feedback with ratings, and behavioral rationales for ratings given by experts. The major difference is the lack of discussion following the presentation of true scores. McIntyre et al. (1984) were the first to assess the effectiveness of FOR with respect to accuracy. They found FOR to result in greater distance and correlation accuracy than RET and a no-training group. They also found that a combined training program, training in FOR and RET, resulted in greater distance and correlation accuracy than no training and RET; however, this group was not more accurate than the FOR group alone. They conclude that perhaps adding RET is not cost effective. Perhaps so, when evaluating RET with the criteria of differential or correlational accuracy. In a later study (Athey & McIntyre, 1987) they compared FOR, INFO (verbal specification of dimensions), and no training. The FOR trained group remembered more of the training content, and provided more distance accurate ratings

than the Info and no training groups. However, there was no significant effect of training on correlation accuracy.

The conclusions concerning RAT effectiveness are not convincing since those trained in RAT perform no more accurately than control groups. Similarly, FOR training research is also not especially persuasive since its resultant accuracy has been mixed. It is likely, however, that the effectiveness of these training efforts have been underestimated. Perhaps it is not the training which is problematic but the evaluation indices used to assess the training. Since FOR training is designed to "tune in" raters to a common frame of reference, we need to assess whether trainees' frames are so "tuned". Similarly, RET is designed to train raters to change rating distributions, and therefore we need to examine the change in distributions.

Alpha, Beta, and Gamma Change

Many organizational development interventions (OD) are evaluated by means of self report outcomes. Prior to 1976, a change in *self-report* in the OD literature,

was thought to be an indication of actual change. A lack of self-reported change was interpreted as intervention failure. Golembiewski, Billingsley, and Yeager's (1976) work changed this conceptualization. They highlighted the fact that OD's unitary criterion of self-reported change is inappropriate and misleading and subsequently leads to diminished or misguided applied research. Prior to their work, most organizational development evaluations recognized only what they referred to as alpha change, actual observed change (i.e., an actual numerical change in self-report measure from time 1 to time 2). Their work demonstrates the utility and practical significance of recognizing other types of change when using self-report measurements--what they labelled beta and gamma change.

Alpha change "involves a variation in the level of some existential state, given a constantly calibrated measurement instrument related to a constant conceptual domain" (Golembieski et al., 1976, p.134). The most noted example of alpha change is of a parent taking a child to a shoe store. The parent is interested in alpha change. The constant conceptual domain, or frame

of reference is the size of the child's feet; the stable measurement instrument, or indicators which are more or less constant, are the instruments such as tape measures or rulers used to assess foot size. If the conceptual domain or the measurement instrument were not stable, it would be senseless to inquire about the change in the child's foot size. However, when the conceptual domain and measuring instrument are stable, foot size comparisons over time are meaningful.

Beta Change "involves a variation in the level of some existential state, complicated by the fact that some intervals of the measurement continuum associated with a constant conceptual domain have been recalibrated" (Golembieski et al., 1976, p.134). In other words, "the respondent alters his or her subjective metric or scale " (Millsap & Hartog, 1988, p.574). After an OD intervention, subjects may make different estimates of reality by changing their subjective intervals of the phenomenon under investigation. The respondent's yardstick for the variable shifts or stretches such that benchmarks on the scale do not remain constant (Tennis, 1989). For example, in an appraisal situation, a change in rating

may result from a change in the supervisor's evaluative standards of ratee quality rather than an actual change in ratee behavior (Millsap & Hartog, 1988). In this example, it would not be meaningful to consider alpha change, because the supervisor's perceived scaling of ratee quality is different from the pre-test scale intervals. The assessment of beta change, is however, important in its own right--especially for those interventions in which a perceived change in measurement intervals is an intended effect. For instance, beta change is an intended effect of some forms of performance appraisal training. Many performance appraisal paradigms include a presentation of behavioral anchors and information on scale calibration. In this instance, a change in the evaluative rating of ratee quality as a function of changed standards is an intended intervention effect.

Gamma change is "a shift in the meaning or conceptualization of the construct being measured" (Millsap & Hartog, 1988, p.574). In other words, it is a major change in perspective or frame of reference of the phenomenon under study (Golembieski et al., 1976). For instance, a post-intervention measure can reflect

an entirely different view of behavior being rated (Millsap & Hartog, 1988). In the rating example, pre-intervention management effectiveness may have been viewed as walking the floor and directing people at work, whereas after training, management effectiveness may be viewed as coordinating activities, delegating, planning, negotiating and coaching. If there is evidence of gamma change, beta and alpha comparisons are not meaningful since the construct under investigation has changed. However, like beta change, some OD interventions (e.g., FOR performance appraisal training), have gamma change as an intended effect.

Alpha, Beta, and Gamma Congruence

The area of performance appraisal training appears plagued with the problem of seemingly inadequate criteria inhibiting our training advancements. Although the performance appraisal accuracy research has not been concerned with the measurement of *intra-individual change* it has been concerned with *inter-individual congruence* (i.e.,the congruence between "experts" and trainees). More specifically, this paper proposes that appraisal

research has been inappropriately concerned with alpha level assessments of training effectiveness. This paper also proposes the new constructs of beta and gamma "congruence" and the measurement of these constructs for training evaluations. The constructs of alpha, beta, and gamma congruence are based on and are analogous to the intra-individual constructs of alpha, beta, gamma change. Alpha, beta, gamma as applied to inter-individual congruence is defined below.

Alpha congruence will be defined here as the actual observed score congruence between trainee and expert scores (i.e., any of the several operational measures of accuracy as traditionally defined). To date, we have assessed performance appraisal training effectiveness vis-a-vis accuracy; accuracy is however, analogous to an alpha level assessment. Similar to the notion discussed above that alpha change should not be assessed if gamma or beta change has occurred, alpha congruence assessments presume gamma and beta congruence; that is, experts and trainees have a similarly perceived calibrated measurement instrument and the same conceptual domain. In a performance evaluation situation, the assessment of alpha

congruence would only be appropriate if trainees and experts have the same perceived calibrations of the measuring instrument and the same conceptual domain. These parameters are, however, only known to be present when objective job performance such as a number of words per minute typed is evaluated. In this example, trainee--expert scale calibrations and conceptual domains would likely be the same (number of words correctly typed and typing ability). However, when performance evaluations are subjective and judgmental, like a majority of appraisal situations, we can not presume that trainee-expert conceptual domains and scale calibrations are identical. We need to go beyond the assessment of alpha congruence.

Beta congruence refers to correspondence of measurement intervals (i.e., the measurement scale) between trainees and experts. Gamma congruence is correspondence in the meaning or conceptualization of the construct being measured. Similar to the notion in the change literature, beta congruence investigation is only meaningful if gamma congruence is present. If gamma congruence is not present, an assessment of beta congruence is not meaningful since the frames of

reference are different for the experts and trainees.

In an appraisal context, the rater must subjectively evaluate a variety of dimensions (e.g., when evaluating teachers we must judge dimensions such as preparation, interest, and examples). A common expert--trainee metric and/or conceptual domain is not necessarily present and can not be assumed. Even in those situations in which alpha level accuracy is apparently present, we need to know whether the experts and trainees share similar scale calibrations and construct conceptualizations. If beta and gamma congruence are absent, the apparent alpha accuracy is a spurious expert--trainee correspondence. In addition, beta and gamma congruence may be present without alpha congruence. The lack of alpha congruence may be due to a variety of causes (e.g., differences in rating experience, halo from former appraisals, etc.). Assessments of beta and gamma congruence would provide important information about the utility of our performance appraisal training efforts.

When raters are trained in RET they are being trained to change their rating distributions, and such changes are inappropriately measured as traditional

accuracy. Similarly, FOR training is also not directly assessed vis-a-vis current accuracy assessments. FOR training is designed to "tune" in raters to a common frame of reference so that to assess its effectiveness we need to establish whether the trainees and experts share a common conceptual domain. Again, our current accuracy assessments do not address this issue.

In sum, there are two major problems with the current performance appraisal criterion of accuracy--which is analogous to an "alpha congruence" assessment: 1) FOR training should induce gamma congruence, and RET training should induce beta congruence, however only alpha congruence has been examined; and 2) the measurement of accuracy (alpha congruence) is not meaningful unless the trainees and experts share the same conceptualization of the evaluation constructs and metric or scale perceptions. Thus, as was true in the OD literature, an alpha level concept of change or in this case congruence, is less than adequate. The aim however of achieving beta and gamma congruence is to ultimately achieve alpha congruence.

Evaluating Alpha, Beta, and Gamma Congruence

Alpha congruence can be evaluated, assuming beta and gamma congruence, using the accuracy assessments currently utilized in the performance appraisal literature as discussed above. Beta and gamma congruence may be analyzed using adaptations of the procedures described below for assessing beta and gamma change.

There is much less consensus regarding the proper method for the assessment of beta change than gamma change. Golembiewski et al. (1976), the individuals responsible for coining the constructs alpha, beta, gamma change do not offer a methodology for beta change assessment. Moreover, they provide no absolute criteria for differentiating beta and gamma change; their technique described below is best used for rejecting gamma change rather than assessing beta change. However, since Golembiewski et al.'s (1976) work several methods have been proposed. Many of these methods require additional data collection of participant perceptions of ideal or retrospective organizational conditions to determine change; (Armenakis & Bedeian, 1982; Armenakis & Zmud, 1979;

Bedeian, Armenakis, & Gibson, 1980; Terborg, Howard, & Maxwell, 1980; Terborg, Maxwell, & Howard, 1982; Zmud & Armenakis, 1978) as such, they are not relevant for measuring beta congruence. Additional data methods include: asking about "ideal" conditions in addition to measuring actual organizational conditions pre- and post-intervention (Armenakis & Bedeian, 1982; Armenakis & Zmud, 1979; Bedeian, Armenakis, & Gibson, 1980; Zmud & Armenakis, 1978), and 2) collecting retrospective "then" measures (Terborg et al., 1982; Terborg et al., 1980). Both these procedures focus on intra-individual change perception measurement and thus are not relevant for examining trainee--expert congruence.

Schmitt (1982) and Millsap and Hartog (1988) provide statistical models for beta change assessment. Schmitt (1982) used a latent variable model of pre-test and post-test scores of perceived motivation before and after seeking employment. Schmitt (1982) proposed that differences between pre- and post-test latent variable variances indicates that scale units are different and beta change has taken place. Millsap and Hartog (1988) propose an approach based on Schmitt's (1982) work but instead focus on regressing post-test latent variables

on pre-test latent variables. They assume that beta change in the experimental group will alter 1) the linearity of the regression and 2) the size of the regression coefficient in comparison to the control group. This method of Beta change assessment however can only be detected by examining pre-post regressions of both treatment and control groups. Pre-post congruence does not apply to the newly proposed beta congruence construct; again, beta congruence is an inter-individual assessment not an intra-individual assessment. Thus to analyze beta congruence the procedures employed by Schmitt (1982) will be employed. In other words, beta congruence will be examined vis-a-vis latent variable variances. If these variances are different for trained and expert groups, it would indicate that scale units are different and beta congruence is absent. If they are not different beta congruence would be evident.

The predominant definition of gamma change is change in factor structure from pre- to post-test; however, techniques used to measure this are varied. Specifically, Golembiewski et al. (1976) proposed using a rotational technique (Ahmavaara, 1954) to rotate pre-

and post- test factor matrices to maximum similarity. Summary indices of similarity are then applied (i.e., coefficient of congruence) (Zmud & Armenakis, 1978). In this paradigm, high factor congruence indicates a lack of gamma change, low congruence indicates gamma change, and moderate congruence is thought to be capturing beta change. Schmitt (1982) proposed studying factor structure change using confirmatory factor analysis adapted from Werts, Rock, Linn and Joreskog (1977) rather than the above rotational procedures. Millsap and Hartog (1988) similarly apply a confirmatory factor analysis technique and operationalize gamma change as any change in the factor pattern matrices from pre to post. Inter-individual gamma congruence can also be analyzed vis-a- vis factor structures. When trainees and experts exhibit similar factor structures as assessed by using confirmatory factor analysis, gamma congruence is present. When factor structures are significantly different, gamma congruence is absent.

In sum, the methodology used to assess *intra-individual* beta and gamma change can be modified to accomodate the measurement of *inter-individual* beta and

gamma congruence. Beta congruence is operationally defined as congruence of latent variable variances between experts and trainees. Gamma congruence is operationally defined as similarity of factor structures between trainees and experts.

The Validity of Expert True Score Estimates

In order to assess accuracy, true score estimates of performance are necessary. Researchers of appraisal rating accuracy identify true scores through the utilization of experts. Several populations have been considered expert: Industrial-Organizational Psychology graduate students (Athey & McIntyre, 1987; Borman, 1977; McIntyre, Smith & Hassett, 1984; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982; Murphy & Balzer, 1986; Pulakos, 1985), Counseling Psychology students (Borman, 1977; Pulakos, 1985), Graduate students (concentration unknown) (Hahn & Dipboye, 1988), Personnel department employees (Hahn & Dipboye, 1988), practicing Industrial/Organizational Psychologists (Borman, 1977; Pulakos, 1985), and undergraduates (Bernardin & Pence, 1980). Industrial/Organizational Psychology graduate students

appear to be the most popular expert population. This is a logically appealing expert group since they are familiar with rating error, the job in question (usually a teacher or manager), and many times the stimulus materials themselves; they also are generally readily accessible as research participants.

There are two primary methods by which true score estimates have been established. The first is based on Borman's (1977) notion that given enhanced opportunities to examine videotapes or scripts, the average rating computed over a number of expert judges provides a true score measure of the ratee's performance. This enhanced opportunity to study stimulus materials and averaging of ratings have been used by many researchers (Bernardin & Pence, 1980; Borman, 1977; Hahn & Dipboye, 1988; Murphy, Garcia, Kerkar,Martin, & Balzer, 1982; Murphy & Balzer, 1986; Pulakos, 1985). An alternative has been employed by McIntyre and his colleagues (McIntrye et al., 1984; Athey & McIntyre, 1987). The "experts" studied the videotapes and scored the ratees' performance, they then discussed and compared ratings until consensus was reached. Consensus ratings were used as true score

estimates. This procedure has only been utilized by McIntyre and his colleagues (McIntrye et al., 1984; Athey & McIntyre, 1987). They also are the only authors who investigated the accuracy of FOR training; consensus and discussion was necessary to establish the frame of reference for their training program intervention. Problematically, the consensus procedure for establishing true scores does not include the quantitative assessment of whether experts disagreed and the extent of agreement (Sulsky & Balzer, 1988). As Sulsky and Balzer (1988) highlight, the "true" rating could be a product of large disagreements among experts. In addition, the consensus procedure is subject to all the problems associated with group decision making (i.e., group-think, conformity, winning argument may take dominance over arriving at the best solution, etc. (Yukl, 1981)).

Since a majority of performance appraisal training accuracy studies use expert true scores, it is important to question their adequacy. One indirect form of validation is the logic of the true score estimate generation procedure (Smither, Barry, & Reilly, 1989). The process of averaging ratings over a

number of experts with multiple opportunities to view videotaped performance has face validity. More direct forms of validation involve examining convergent and discriminant validity (Borman, 1978; Murphy et al., 1982; Murphy & Balzer, 1986), or inter-rater agreement (Borman,1977; Pulakos,1984). Convergent validity has been examined by analyzing the intraclass index for the ratee main effect in a Rater X Performance Dimension design (Borman, 1978; Murphy et al., 1982; Murphy & Balzer, 1986). The intraclass index for Ratee X Dimension has been used to measure discriminant validity (Borman, 1978; Murphy et al., 1982; Murphy & Balzer, 1986) and inter-rater reliability has been assessed by computing an intraclass correlation coefficient for each job and each dimension (Borman, 1977).

Most recently, Smither et al. (1989) conducted the first published investigation of the relevance of expert true score estimates. They compared objective true scores defined in terms of objective worker output with mean expert ratings. Experts were graduate students in Industrial/ Organizational Psychology. Students were given ample opportunity to observe

videotaped performances of customer sales representatives for a telecommunications company. True performance was objectively defined as the number of sales, authorizing and processing paperwork, and resolving customer account problems on a computer. They compared objective true scores with non-expert undergraduate ratings. They found expert raters to be more accurate than non-expert raters (differential and correlation accuracy). In order to evaluate the accuracy of mean expert ratings compared to objective true scores, objective true scores (known performance levels, e.g., resolved 3 customer account problems) were converted into an evaluation scale (poorest, average, best). Accuracy indices computed using objective true scores were highly correlated with mean expert ratings (the correlation between mean expert ratings and objective true scores was nearly perfect (r=.99). Although these findings are encouraging, systematic differences in absolute level(s) of ratings are not reported.

In sum, convergent and discriminant validity, or inter-rater reliability may be used to assess expert ratings. In addition, we can be fairly confident with

these ratings since they have been found to be highly correlated with objective true scores and have high face validity.

Problem Statement

Training effectiveness has recently been assessed by examining accuracy ("alpha congruence"), the correspondence of observed performance ratings with true score estimates provided by expert raters. Alpha congruence is not meaningful unless there is also evidence of beta and gamma congruence. In addition, measures of accuracy defined this way may not reveal the effects of performance appraisal training. For example, FOR is designed to tune raters to a common frame of reference so that behaviors can be similarly assessed by different raters. Researchers have not found that FOR training consistently leads to greater accuracy. This does not, however indicate that greater congruence in the interpretation of rating dimensions was not achieved. The failure to increase alpha accuracy could be due to other reasons. The effectiveness of FOR can more directly be assessed as gamma congruence, the congruence between the trained

and expert frame of reference, applying the procedures used to assess gamma change. Once a common frame of reference has been established we can also assess whether trained raters change their rating scale calibration toward that of the experts by assessing beta congruence, applying the procedures used to assess beta change (Millsap & Hartog, 1988).

It is hypothesized that:

1) Those trained in FOR will exhibit trainee standards (factor structure) that are congruent with those of expert raters (gamma congruence). As described above, this in fact is the intent of the FOR training paradigm. FOR training instructs trainees to use expert based constructs for evaluating behaviors. Evaluations to date have only addressed alpha level assessments which do not necessarily indicate appropriate utilization of rating frames.

2) Those trained in FOR will not demonstrate rating scale calibration congruence with experts' (no beta congruence). This is because FOR does not train raters

to alter their scale calibrations, instead it instructs trainees to use expert based constructs for evaluating behavior.

3) Those trained in FOR combined with RET will display expert congruent standards (factor structure) and scale calibration (gamma and beta congruence). The rationale for similar trainee expert standards is described in the first hypothesis. RET training will have the added value of beta congruence since its focus is on rating distributions. When experts and trainees share a common frame of reference, RET will result in rating scale calibration congruence with experts. Note: RET training alone will not be assessed since without a proper frame of reference (gamma congruence), beta comparisons are not meaningful.

CHAPTER II

METHOD

Subjects

Two-hundred thirty six (236) undergraduate psychology students from City University were recruited to serve as trainees.

All participants signed an informed consent form (Appendix A).

Stimulus Materials

Two videotaped lectures, one on self-fulfilling prophecy and another on crowding and stress, developed by Murphy and his colleagues (Murphy, Garcia, Kerkar, Martin & Balzer, 1982) served as the practice tape and rating stimulus, respectively. Murphy et al. (1982) developed eight videotaped lectures using "a common outline for the lectures in each of the two content areas but the thoroughness and organization of each lecture and the clarity of the responses to questions were varied between and within lectures" (p.322). The two lectures selected received an overall average

rating by expert raters in both Murphy et al.'s (1982) and McIntyre et al.'s (1984) work. Each lecture lasts approximately six minutes.

Rating Scale

A 16-item rating scale (see Appendix B) based on McIntyre et al.'s (1984) 12-item rating scale and Murphy et al.'s (1982) 12-item scale were used to assess lecture performance. The scale was constructed to measure three dimensions of teaching effectiveness : 1) physical aspects of presenter (i.e., eye contact, tone of voice, facial expression); 2) aspects of the presented material (i.e., number of examples, helpfulness of examples, clarity of answers) 3) organization of content (i.e., transitions between subtopics, summary statements). Background information and training knowledge measures were also gathered (Appendix C).

True Score Estimates

Prior to viewing the crowding and stress lecture, experts were shown the rating scale. After their ratings of the crowding and stress lecture were

completed, they read the established frames of reference from McIntyre et al. (1984). McIntyre et al.'s (1984) frames of reference provided these experts with other experts' ratings and justification of each rating. Subsequently, the lecture on self-fulfilling prophecy was evaluated. Experts were given as much time as they liked to study and rate the videotapes (each tape was viewed at least twice). Expert group instructions are presented in Appendix D.

Experts

Expert ratings were obtained from seventy-four (74) expert raters: nine (9) Ph.D. Industrial/Organizational Psychologists and sixty-five (65) Industrial/Organizational Psychology doctoral students from Baruch College, Stevens Institute of Technology, New York University and Hofstra University. Mean ratings were used as true score estimates of lecturer performance to assess alpha congruence. Expert raw scores were used to assess beta and gamma congruence.

Procedure

Twelve intact classes were randomly assigned to one

of three conditions: Frame of Reference Training (N=76), Frame of Reference and Rater Error Training (N=77), or no training (N=83).

Rater Training

The FOR training employed by McIntyre et al. (1984) was followed. This included reading aloud the 16 dimensions from the rating scale and encouraging questions and discussion. Frame of Reference Training included a practice rating with feedback of true scores and explanation of the rationale for true scores by pointing out behaviors attended to by the experts. FOR training lasted approximately 1/2 hour (See Appendix E).

Frame of Reference Training combined with Rater Error Training lasted approximately 50 minutes and included a presentation of halo, leniency, and range restriction errors (See Appendix F). It also included reading aloud the 16 dimensions from the rating scale and encouraging questions and discussion. Participants also engaged in a practice rating and received true score feedback. RET/FOR was identical to FOR in process (i.e., lecture, practice, feedback) but the

lecture and feedback content focused on both error and expert frames.

The no training group received a lecture on crowding and stress to supplement material presented in the videotaped lecture (See Appendix G).

Analyses

Ratings obtained from subjects were analyzed for gamma, beta, and alpha congruence. The LISREL VI and VII structural equations computer program (Joreskog & Sorbom, 1985; 1988) were used to assess beta and gamma congruence. Item-level data for each group served as input into the LISREL analyses. Covariance matrices were then analyzed to assess factor structure (gamma) and factor covariance congruence (beta) across groups.

Gamma Congruence

A confirmatory factor analytic procedure developed by Werts et al. (1977) and adapted by Schmitt (1982) was used to assess congruence in factor structure. Factor structure congruence between groups was compared by imposing equality constraints on factor pattern matrices (lambda). Imposing equality restricts

the factor pattern matrix to be invariant, i.e., each free element and starting value in the trained groups were set to be equal to the expert group. LISREL provides a chi-square that can be used to compare a constrained model (i.e., factor pattern equality) with another model that includes fewer constraints (i.e., no factor pattern equality). An increase in chi square compared to a less constrained model indicates factor incongruence.

Beta Congruence

Congruence in factor scales was compared when factor patterns were equal by additionally imposing invariance constraints on the factor covariance matrix (phi) (Schmitt, 1982). Imposing equality on phi restricts the factor covariance matrix to be invariant between the expert group and the trained group being analyzed.

Alpha Congruence

Traditional measures of accuracy, elevation (E) (Cronbach, 1955) and distance accuracy (DA) (Athey & McIntyre, 1987; McIntyre et al., 1984) were computed.

Elevation and accuracy were utilized since computations do not require ratings of multiple ratees. These accuracies may be expressed by the following formulae:

1) $E^2 = (\bar{x}_{..} - \bar{t}_{..})^2$, where $x_{..}$ and $t_{..}$= mean rating and mean true score, over all ratees and items.

2) $$DA_k = \frac{\sum_{j=1}^{n} \left(\sum_{j=1}^{d} |t_{ij} - r_{ijk}| \right) / d}{n}$$

where k refers to the k^{th} rater; n is the number of ratees, d is the number of dimensions, r refers to the subject rating; and t refers to true scores (Sulsky & Balzer, 1988).

Hypothesis Testing

The hypotheses tests examined a series of nested models to determine: a) the number of factors that best describe the data (i.e., two, three, or four factors), and, b) if gamma and beta congruence between experts and trainees was evidenced. Models were examined in succession with constraints added during each step. This hypothesis testing strategy is applicable whenever

a more restricted model can be created by imposing constraints on a more basic model (Hayduk, 1987).

The following series of nested hypotheses were tested to assess gamma and beta congruence: unequal factor patterns and factor covariances, equal factor patterns (gamma congruence), equal factor patterns and factor covariances (gamma and beta congruence). The series of tests examined whether added restrictions (lambda invariance, phi invariance, and no common factors) based on the expert group and imposed on the trained groups significantly reduced the fit of the model. The hierarchy of models from least to most constrained is presented in Table 1. Differences between less and more constrained models were evaluated using Chi-square (χ^2). The evaluation of differences in Chi-square between nested models is further described below.

Fit Indices

The following provides a brief foundation for the interpretation of each fit index used. Fit indices provide an indication of the magnitude of lack of fit of the tested model. It is important to bear in mind

Table 1

Hierarchy of Models for Two Group Case

(least constrained to most constrained)

Group 1		Group 2	
Λ_1, Φ_1 Θ_1	$\Lambda_1 \neq \Lambda_2$ $\Phi_1 \neq \Phi_2$	Λ_2, Φ_2 Θ_2	Same Number of Factors per Group
Λ, Φ_1 Θ_1	$\Lambda_1 = \Lambda_2 = \Lambda$ $\Phi_1 \neq \Phi_2$	Λ, Φ_2 Θ_2	Gamma Congruence No Beta Congruence
Λ, Φ Θ_1	$\Lambda_1 = \Lambda_2 = \Lambda$ $\Phi_1 = \Phi_2 = \Phi$	Λ, Φ Θ_2	Gamma Congruence Beta Congruence
$\Lambda = 0$ $\Phi = 0$ Θ_1	$\Lambda_1 = \Lambda_2 = 0$ $\Phi_1 = \Phi_2 = 0$	$\Lambda = 0$ $\Phi = 0$ Θ_2	Null

Λ = Lambda; Φ = Phi; Θ = Theta

that these indices are not absolute, but merely rules of thumb, and should be interpreted accordingly. Consulting all the indices gives the most accurate picture of the situation.

Chi-square

The most commonly employed fit index is chi-square. Interpretation of the Chi-square statistic should be as "lack of fit" (Muliak et al., 1989). As discrepancies between model and data increase, so does the value of Chi-square. The degrees of freedom serves as a standard from which to judge whether Chi-square is large or small. The Chi-square measure is sensitive to both sample size and departures from multivariate normality of the observed variables (Joreskog & Sorbom, 1986). "Large sample sizes and departures from normality tend to increase Chi-square over and above what can be expected due to specification error in the model" (Joreskog & Sorbom, 1986, p.I.39).

When nested models are examined the differences between Chi-square (χ^2_{diff}) may be analyzed, $\chi^2_{diff} = \chi^{2constrained} - \chi^{2less\ constrained}$. The difference between two Chi-squares is distributed as Chi-square

with degrees of freedom equal to the difference in degrees of freedom for the two models. Differences in Chi-square indicate whether additional constraints reduce the tested model's ability to fit the data.

NFI

The normed fit index (NFI), otherwise known as the Bentler-Bonet Index, uses the Chi-square distribution. NFI= $(\chi^2_k - \chi^2_o) / \chi^2_k$, where k and o refer to the model in question and the null model respectively. Although values above .90 are commonly considered acceptable a cut-off of .80 was imposed.

LISREL'S GFI

LISREL's goodness of fit index (GFI) measures "the relative amount of variances and covariances jointly accounted for by the model" (Joreskog & Sorbom, 1989, p.I.40). In general, values of .90 or better are considered to be indicative of good fit, however, a cut-off of .80 was imposed. GFI is bounded by 0 and 1.

LISREL'S RMSR

The root mean square residual (RMSR) is also

provided in the output from the LISREL program. This index is the square root of the mean squared residual obtained from the difference between the observed and reproduced correlation matrices (Joreskog & Sorbom, 1989, p.I.40). Higher values imply higher error and lower fit. To decide whether there is a lack of fit RMSR is compared to the average size of the elements in the covariance matrix. Based on the covariance matrices a cut-off of .15 was imposed. Covariance matrices are presented in Appendix H.

Problems with Maximum Likelihood Estimation

When using maximum likelihood estimation in sample sizes less than 100, estimated unique variances are frequently negative (Hayduk,1987). When this situation is encountered, the solution is improper, and termed a Heywood case (Van Driel, 1978). Also with small sample sizes, the problem of phi being not positive definite is more commonly confronted. Phi not positive definite is an indication of poor model fit, often indicating that there are too many factors. These problems are relevant to the present study since samples are fairly small ($N < 100$).

CHAPTER III

RESULTS

Background, Appraisal Experience and Task Perception Differences between Experts and Experimental Groups

As expected, there are significant differences between expert and experimental group demographics, appraisal experience, and task perceptions. Demographic information for experts and trainees is presented in Table 2. Experts were older than trainees [($\underline{M}$ = 27.88) ($\underline{M}$ = 23.18), $\underline{t}(307) = -5.62$, $\underline{p}<.001$], more educated [$\chi^2(8, \underline{N} = 309) = 281.41$, $\underline{p}<.001$], and were more likely to be currently working [$\chi^2(2, \underline{N} = 309) = 20.92$, $\underline{p}<.001$].

With respect to appraisal experience, more experts than trainees had heard of Rater Error Training [$\chi^2(1, \underline{N} = 306) = 144.19$, $\underline{p}<.001$] and Frame of Reference training [$\chi^2(1, \underline{N} = 306) = 62.07$, $\underline{p}<.001$]. In addition, more experts had rated or viewed the videotapes before [χ^2 $(1, \underline{N} = 306) = 5.67$, $\underline{p}<.05$]. The experts also reported having more experience with

Table 2

Demographics by Group

	FOR/RET	FOR	Comparison	Expert
Age (mean)	23.63	24.03	21.98	27.88
Sex				
Male	36%(28)	47%(36)	40%(33)	45%(33)
Female	64%(49)	53%(40)	60%(50)	55%(41)
Highest Educational Level				
High School	3%(2)	0%(0)	2%(2)	0%(0)
Freshman	17%(13)	17%(13)	13%(11)	0%(0)
Sophomore	12%(9)	26%(20)	25%(21)	0%(0)
Junior	3%(2)	0%(0)	2%(2)	0%(0)
Senior	50%(35)	26%(20)	19%(16)	0%(0)
BA,BS, BBA	1%(1)	5%(4)	4%(3)	0%(0)
Some Graduate	1%(1)	3%(2)	2%(2)	28%(21)
MA, MS, MBA	1%(1)	0%(0)	0%(0)	60%(44)
Ph.D.	0%(0)	0%(0)	0%(0)	12%(9)
Working Situation				
Student	22%(17)	15%(11)	45%(37)	19%(14)
Working and attending school	78%(60)	86%(65)	55%(46)	73%(54)
Working and not attending school	0%(0)	0%(0)	0%(0)	8%(6)

having their performance evaluated than trained groups [(M = 3.58) (M = 2.90), t(304) = -3.55, p<.001] and being more familiar with the job of teaching [(M = 3.52) (M=2.32), t(303) = -6.23, p<.001] than trained groups . Appraisal experience information is presented in Table 3.

Experts felt more confident than the trainees that their ratings were accurate [(M = 3.42) (M = 3.25) t(303) = -1.90, p<.05]. Unlike the undergraduate groups, experts did not anticipate that training would help them in the future [(M = 3.33) (M = 2.60) t(301) = 4.56, p<.001]. Expert task perceptions are presented in Table 4.

The demographic, appraisal experience, and task perception differences between the expert and experimental groups affirm that the "experts" seem to indeed be "expert".

Background, Appraisal Experience and Task Perception Differences between Experimental Groups

Demographics for the experimental groups are also presented in Table 2. There are few significant differences among the experimental groups. The

Table 3

Appraisal Experience by Group

	FOR/RET	FOR	Comparison	Expert
Experience evaluating job performance.[1]	.86	1.28	1.16	1.41
Experience evaluating teachers.[1]	3.39	3.12	3.40	3.20
Experience having your own performance evaluated.[1]	2.68	2.59	3.10	3.58
Heard of RET.[2]	25%(19)+	9%(7)	9%(7)	89%(66)**
Heard of FOR.[2]	16%(12)	25%(19)+	10%(8)	64%(47)*
Rated or viewed video before.[2]	5%(4)	1%(1)	8%(6)	14%(10)*
Familiar with the job of teaching.[3]	2.23	2.36	2.38	3.52**

NOTE:

[1] Ratings are based on the following 5 point scale: 0) None ,1) less than 1, 2) more than 1 less than 3, 3) more than3 less than 5, 4) more than 5 less than 10, 5) 10 or more. The presented numbers are means.

[2] Ratings are based on the Yes/No answers, the presented numbers represent the percentage whom responded yes.

[3] Ratings are based on the following 5 point scale: 0) Not at all ,1) to a very small extent, 2) to a small extent, 3) to a moderate extent, 4) to a great extent, 5) to a very great extent. The presented numbers are means.

\+ **p<.05** Differences between experimental groups.

++ **p<.01** Differences between experimental groups.

* **p<.05** Expert group differed from experimental groups.

** **p<.01** Expert group differed from experimental groups.

Table 4
Task Perceptions by Group

	FOR/RET	FOR	Comparison	Expert
I am confident that my ratings are accurate.	*3.09*	3.17	*3.47+*	*3.42**
The lecturer would consider my ratings fair.	*2.95*	3.32	*3.51++*	3.35
I feel comfortable rating teachers.	3.93	3.84	3.72	3.64
The training was interesting.	3.14	3.24	2.87	3.08
Training helped me make more accurate evaluations.	3.09	3.30	2.88	2.96
I anticipate that training will help me in the future.	3.39	3.41	3.19	*2.60***

NOTE:
Ratings are based on the following 5 point scale: 0) Not at all ,1) to a very small extent, 2) to a small extent, 3) to a moderate extent, 4) to a great extent, 5) to a very great extent. The presented numbers are means.

\+ **p<.05** Experimental group differed from other experimental group. **Bold** print identifies where the differences are.

++ **p<.01** Experimental group differed from other experimental group. **Bold** print identifies where the differences are.

* **p<.05** Expert group differed from experimental groups.

** **p<.01** Expert group differed from experimental groups.

differences that do exist are as follows: in the comparison group there are fewer students who are working [$\chi^2(2, N = 236) = 19.72, p<.001$], and the educational level is highest in the FRET group [$\chi^2(14, N = 236) = 24.74 , p<.05$].

Experience with appraisal is reported in Table 3. More trainees in the combined training group (frame of reference and rater error) have heard of RET before [$\chi^2 (2, N = 233) = 10.46, p<.01$], and more trainees in the Frame of Reference group have heard of FOR before [$\chi^2 (2, N = 232) = 6.35, p<.05$].

Task perceptions are reported in Table 4. The comparison group was more confident that their ratings were accurate than was the FRET group [($M = 3.47$) ($M = 3.09$), $F(2, 229) = 4.25, p<.05$]. They also felt more strongly that the lecturer would consider their ratings fair than did FRET group [($M = 3.51$) ($M = 2.95$), $F(2, 226) = 6.83, p<.01$].

Scale Reliability

Alpha reliability on the combined experimental group data (i.e., FOR/RET, FOR, and Comparison groups) was assessed for the lecturer evaluation scale.

Reliability for the scale was found to be satisfactory (α =.92).

Manipulation Check

The manipulation checks conducted examined frame of reference knowledge for the groups trained in FOR and error knowledge for those trained in RET. Expert groups also completed a knowledge test on frames of reference and errors. The manipulation check showed that the two experimental groups trained in FOR (Frame of reference training and frame of reference training combined with rater error training) did not perform any differently from each other in a knowledge test regarding frames of reference. However these two trained groups did significantly differ from the experts on the following manipulation check questions: 1) True or false, to evaluate "interest in the topic" you should look at whether he provided a summary statement: Experts = 82% correct, Trained = 65% correct, [$\chi^2(1, N = 215) = 5.76, p<.05$], 2) True or false, to evaluate "he used clear examples to explain abstract ideas" you should look at whether the lecture was smooth: Experts = 93% correct, Trained = 69%

correct, [$\chi^2(1, N = 221) = 14.83, p<.001]$, 3) True or false, to evaluate "he followed a logical sequence of thought in his lecture" you should look at whether answers were clear: Experts = 85% correct, Trained = 55% correct, [$\chi^2(1, N = 220) = 17.79, p<.001]$, 4) True or false, to evaluate "he was well prepared" you should look at facial expressions: Experts = 90% correct, Trained = 69% correct, [$\chi^2(1, N = 221) = 11.14, p<.001]$.

No significant differences were found in knowledge of rating errors between the experimental group trained in error and the experts. Both groups correctly identified halo, restriction of range, severity and leniency.

Factor Extraction

Chi-square was used to compare a constrained model (i.e., no common factors) with another model that includes fewer constraints (i.e., two, three and four factor models). These models were evaluated using Chi-square, NFI, GFI and RMSR fit indices described above.

Experts

Expert data were examined to determine the number of factors appropriate for explaining the data. Null (no common factors), two, three, and four factor solutions were examined. For each factor, one item which was the "marker" of that factor (had the highest factor loading, as assessed by preliminary exploratory factor analysis), was forced to load on one factor by having its loading fixed to one (1.00) and all other factor loadings fixed to zero (0). This procedure was conducted to uniquely identify the factor solution. The marker items were: 1) "followed a logical sequence of thought in his lecture", 2) "emphasized important points by raising his voice", 3) "acted relaxed", 4) "was well prepared". For the four-factor solution all four items were used, for the three factor solution the first three were used and for the two factor solution the first two were used. Two items were dropped from further analyses since they exhibited small loadings on the identified factors. The two items dropped were: 1) "he provided relevant answers to the questions" and, 2) "he looked at the class while speaking". In light of the lecturer's performance, it is apparent why the item

regarding the question and answer period did not load on any of the factors. The lecturer's performance during this part of the lecture was extremely poor compared to the other parts of the lecture. The second item dropped, "he looked at the class while speaking", can also be explained. Many participants found eye contact difficult to assess because the videotaped lecture did not have close up pictures of the lecturer. Further analyses revealed that the best fitting solution was the three-factor solution, [$\chi^2(52, \underline{N} = 74) = 63.49$, $p>.05$, NFI = .89, GFI = .894, RMSR = .068]. The four-factor solution revealed that four factors did not fit the data; yielding a "Heywood case". Two factors did not fit as well as three [$\chi^2(64, \underline{N} = 74) = 107.46$, $\underline{p}<.01$, NFI = .81, GFI = .821, RMSR = .115.; $\chi^2_{diff}(12, \underline{N} = 74) = 43.97$, $\underline{p}<.01$]. Fit indices for the null, two and three factor solutions are presented in Table 5.

Factor loadings for each of the items in the best fitting three factor solution are presented in Table 6. Factor 1 was labeled physical aspects of the presenter because it subsumed the following items and frames: he seemed interested in the topic (eye contact, tone of

Table 5

Measurement Model Fit Indices by Group

	INDEX			
MODEL	CHI^2 (df)	NFI	GFI	RMSR
Expert Group				
Null	563.70(91)	--	--	--
3 Factor Model	63.49(52)	.89	.894	.068
2 Factor Model	107.46(64)	.81	.821	.115
FRET Group				
Null	497.70(91)	--	--	--
3 Factor Model	82.88(52)	.83	.873	.080
FOR Group				
Null	619.37(91)	--	--	--
4 Factor Model	55.41(41)	.91	.913	.042
3 Factor Model	86.78(52)	.86	.863	.062
Comparison Group				
Null	928.23(91)	--	--	--
3 Factor Model	138.71(52)	.85	.827	.089

NOTE: CHI^2 = CHI-square value ; NFI = Normed Fit Index; GFI = Goodness of Fit Index; RMSR = Root Mean Square Residual

Table 6

Expert Group (Maximum Likelihood Estimates)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3
Seemed interested in the topic	eye contact tone of voice, facial expression	***.515***	.245	.051
Used examples abstract ideas	how many examples how helpful	.257	***.830***	.285
Presented the lecture smoothly	transitions between topics smooth	.209	.537	***.650***
Integrated the material effectively	good transitions summary statement	.269	***.781***	.340
Followed an outline	include all subtopics, equal time	.170	***1.045***	.246
Followed a logical sequence of thought in his lecture	logical transitions	.000*	***1.000****	.000*
Was well prepared	number of studies examples, responses to questions	.212	***.503***	.450
Acted relaxed	body movement, verbal expressions facial expressions	.000*	.000*	***1.000****
Spoke clearly and distinctly	pronunciation was easy to listen to	.233	.148	***.610***
Spoke with vigor and enthusiasm	tone and volume voice, facial expressions	***1.042***	.058	.146

Table 6

Expert Group (Maximum Likelihood Estimates continued)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3
Emphasized important points by raising his voice	tone and volume voice	*.875*	.031	.168
Voice was animated	tone of voice	*1.000**	.000*	.000*
Examples were presented which were clearly related to central topic	examples useful	.268	*.749*	.040
Used purposeful non-verbal behavior	facial expressions, body movement	*.490*	.223	.382

* *Fixed for identification*

voice, facial expression), he spoke with vigor and enthusiasm (tone and volume of voice, facial expression), emphasized important points by raising his voice (tone and volume of voice), voice was animated (tone of voice), used purposeful non-verbal behavior (facial expressions, body movement). Factor 2 centered on aspects of the presented material and organization of content. Items and frames loading on factor 2 were: used examples to explain abstract ideas (how many examples, how helpful), integrated the material effectively (good transitions, summary statement), followed an outline (included all subtopics, equal time to each subtopic), followed a logical sequence of thought in his lecture (logical transitions), was well prepared (number of studies, examples, and responses to questions), examples were presented which were clearly related to the central topic (examples useful). Factor 3 was termed general platform skills because it included: presented the lecture smoothly (transitions between subtopics smooth), acted relaxed (body movement, verbal expressions, facial expressions), spoke clearly and distinctly (pronounciation, was easy to listen to).

Frame of Reference and Rater Error Training (FRET)

Analyses for the FRET group were performed with the same two items dropped and the same identifiers as in the expert group. The best fitting solution for this group was also a three factor model [$\chi^2(52, N = 77) = 82.88$, $p<.01$, NFI = .83, GFI = .873, RMSR = .080]. Four factors did not fit the data (Heywood case). Two factor models were not calculated since all things being equal the number of factors is positively related to model fit. The fit of the two factor model would be worse than the three and thus would still establish the three factor model as the best fitting model. Fit indices for the null and three factor solutions are presented in Table 5. Factor loadings for the three factor solution are presented in Table 7.

Frame of Reference Training (FOR)

Analyses for the FOR group were performed as above. The best fitting solution for this group was a four-factor model [$\chi^2(41, N = 76) = 55.41$, $p>.05$, NFI = .91, GFI = .913, RMSR = .042]. The three-factor model, although adequate, did not fit as well, [$\chi^2(52, N = 76) = 86.78$, $p>.001$, NFI = .86, GFI = .863, RMSR =

Table 7

Frame of Reference and Rater Error Training (FRET) (Maximum Likelihood Estimates)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3
Seemed interested in the topic	eye contact tone of voice, facial expression	.053	.378	***.548***
Used examples abstract ideas	how many examples how helpful	***1.944***	1.864	-2.022
Presented the lecture smoothly	transitions between topics smooth	-.298	***.603***	.582
Integrated the material effectively	good transitions summary statement	-.748	.745	***.830***
Followed an outline	include all subtopics, equal time	-1.762	.144	***1.663***
Followed a logical sequence of thought in his lecture	logical transitions	.000*	***1.000****	.000*
Was well prepared	number of studies examples, responses to questions	.022	***.565***	.514
Acted relaxed	body movement, verbal expressions facial expressions	.000*	.000*	***1.000****
Spoke clearly and distinctly	pronunciation was easy to listen to	-1.660	-1.452	***3.272***
Spoke with vigor and enthusiasm	tone and volume voice, facial expressions	-1.545	-2.164	***3.961***

Table 7

Frame of Reference and Rater Error Training (FRET) (Maximum Likelihood Estimates continued)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3
Emphasized important points by raising his voice	tone and volume voice	*.873*	-.326	.263
Voice was animated	tone of voice	*1.000**	.000*	.000*
Examples were presented which were clearly related to central topic	examples useful	*2.986*	2.698	-3.789
Used purposeful non-verbal behavior	facial expressions, body movement	1.095	*1.170*	-1.469

* *Fixed for identification*

.062; $\chi^2_{diff}(11, N = 76) = 31.37, p<.01]$. Two-factor models were not calculated since the fit of the two-factor model would be worse than the three and thus would still establish the four-factor model as the best fitting model. The null, three-, and four-factor solution fit indices are presented in Table 5. Factor loadings for the three- and four-factor solutions are presented in Table 8 and Table 9, respectively.

Comparison Group

Analyses for the Comparison group were performed in the same manner as the other groups. The best fitting solution for this group was a three-factor model [$\chi^2(52, N = 83) = 138.71, p>.05$, NFI = .85, GFI = .827, RMSR = .089]. The four- factor model did not fit the data (Heywood case). Two-factor models were again not calculated since the fit of the two factor model would be worse than the three-factor model. The null and three-factor fit indices are presented in Table 5. Factor loadings for the three-factor solution are presented in Table 10.

Table 8

Frame of Reference (FOR) Group (Maximum Likelihood Estimates)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3
Seemed interested in the topic	eye contact tone of voice, facial expression	-.139	.015	***.957***
Used examples abstract ideas	how many examples how helpful	***2.214***	.478	-1.246
Presented the lecture smoothly	transitions between topics smooth	***.911***	.573	-.084
Integrated the material effectively	good transitions summary statement	.345	***.662***	.019
Followed an outline	include all subtopics, equal time	.092	***1.027***	-.223
Followed a logical sequence of thought in his lecture	logical transitions	.000*	***1.000****	***.000****
Was well prepared	number of studies examples, responses to questions	-.374	.645	***.714***
Acted relaxed	body movement, verbal expressions facial expressions	.000*	.000*	***1.000****
Spoke clearly and distinctly	pronunciation was easy to listen to	.351	.191	***.554***
Spoke with vigor and enthusiasm	tone and volume voice, facial expressions	-1.497	-.112	***2.282***
Emphasized important points by raising his voice	tone and volume voice	.282	-.156	***.627***

Table 8

Frame of Reference (FOR) Group (Maximum Likelihood Estimates continued)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3
Voice was animated	tone of voice	*1.000**	.000*	.000*
Examples were presented which were clearly related to central topic	examples useful	*2.231*	.220	-.902
Used purposeful non-verbal behavior	facial expressions, body movement	*1.410*	.160	-.164

* *Fixed for identification*

Table 9

Frame of Reference (FOR) Group (Maximum Likeihood Estimates)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
Seemed interested in the topic	eye contact tone of voice, facial expression	.616	-1.280	.-.460	***1.992***
Used examples abstract ideas	how many examples how helpful	.906	***2.822***	1.283	-3.608
Presented the lecture smoothly	transitions between topics smooth	.569	***1.095***	.499	-.791
Integrated the material effectively	good transitions summary statement	.241	***.825***	.182	-.233
Followed an outline	include all subtopics, equal time	-.096	***1.316***	.110	-.446
Followed a logical sequence of thought in his lecture	logical transitions	.000*	***1.000****	.000*	.000*
Was well prepared	number of studies examples, responses to questions	.000*	.000*	.000*	***1.000****
Acted relaxed	body movement, verbal expressions facial expressions	.000*	.000*	***1.000****	.000*
Spoke clearly and distinctly	pronunciation was easy to listen to	-.185	1.275	***1.674***	-1.659
Spoke with vigor and enthusiasm	tone and volume voice, facial expressions	.034	-2.647	-.591	***3.934***
Emphasized important points by raising his voice	tone and volume voice	.271	-.066	***.699***	.000*

Table 9

Frame of Reference (FOR) Group (Maximum Likeihood Estimates continued)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
Voice was animated	tone of voice	***1.000****	.000*	.000*	.000*
Examples were presented which were clearly related to central topic	examples useful	1.463	***1.838***	.762	-2.505
Used purposeful non-verbal behavior	facial expressions, body movement	***1.000***	.759	.462	-.916

* ***Fixed for identification***

Table 10

Comparison Group (Comp) (Maximum Likelihood Estimates)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3
Seemed interested in the topic	eye contact tone of voice, facial expression	.459	***.560***	.050
Used examples abstract ideas	how many examples how helpful	.245	***1.206***	-.320
Presented the lecture smoothly	transitions between topics smooth	-.443	***.978***	.584
Integrated the material effectively	good transitions summary statement	-.081	***1.233***	.049
Followed an outline	include all subtopics, equal time	.038	***.734***	.127
Followed a logical sequence of thought in his lecture	logical transitions	.000*	***1.000****	.000*
Was well prepared	number of studies examples, responses to questions	.222	.484	***.534***
Acted relaxed	body movement, verbal expressions facial expressions	.000*	.000*	***1.000****
Spoke clearly and distinctly	pronunciation was easy to listen to	.455	-.062	***.577***
Spoke with vigor and enthusiasm	tone and volume voice, facial expressions	***1.099***	-.165	.166

Table 10

Comparison Group (Comp) (Maximum Likelihood Estimates continued)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3
Emphasized important points by raising his voice	tone and volume voice	*1.720*	.408	-.150
Voice was animated	tone of voice	*1.000**	.000*	.000*
Examples were presented which were clearly related to central topic	examples useful	.377	*.946*	-.204
Used purposeful non-verbal behavior	facial expressions, body movement	-.221	*.879*	.393

* *Fixed for identification*

Gamma and Beta Congruence

Factor structure congruence between groups was compared by imposing equality constraints on factor pattern matrices (lambda). The LISREL VI and VII structural equations computer program (Joreskog & Sorbom, 1985; 1988) was used to analyze data from two groups simultaneously with the factor pattern matrix in each of the experimental groups (i.e., FOR/RET, FOR, and Comparison groups) constrained to be equal to the expert group. In a multi-sample or stacked analysis one Chi-square goodness-of-fit measure is provided. Chi-square measures "the fit of all LISREL models in all groups, including all constraints, to the data from all groups" (Joreskog & Sorbom, 1985, p.V.4). Therefore this Chi-square statistic was used to compare the constrained model (i.e., factor pattern equality) with the less constrained model that allowed factor pattern matrices to vary (e.g., lambda was free to vary).

Scale calibration (Beta) congruence was assessed by imposing equality constraints on factor covariances (phi) also using a multi-sample analysis. Beta congruence was only asssessed for the groups which

exhibited gamma congruence. Chi-square was used to compare this most constrained model (i.e., factor pattern equality and factor covariance equality) with the pattern-equal model.

Frame of Reference and Rater Error Training (FRET)

The three-factor model for both the expert and FRET group fit relatively well as shown in Table 11, the chi-square for the stacked three-factor model was adequate [$\chi^2(104, \underline{N} = 151) = 146.37$, $\underline{p}<.01$, NFI = .86]. However, once factor pattern invariance was imposed, the chi-square, NFI, GFI and RMSR revealed a lack of fit, [$\chi^2(137, \underline{N} = 151) = 219.29$, $\underline{p}<.001$, NFI = .79; $\chi^2_{diff}(33, \underline{N} = 151) = 72.92$, $\underline{p}<.001$]. Gamma congruence was absent.

Frame of Reference Training (FOR)

A two-stage analysis was performed to assess gamma congruence in the FOR group. First, a three-factor stacked model (experts-FOR) was assessed then a three-factor expert and four-factor FOR model was analyzed. The three factor model's fit was adequate [$\chi^2(104, \underline{N} = 150) = 150.27$, $\underline{p}<.01$, NFI = .87], but not as good as

Table 11

Measurement Models Fit Indices For Stacked Runs: Frame of Reference and Rater Error (FRET) and Expert Group

	INDEX					
	CHI^2 (df)	NFI	GFI (Fret)	GFI (Exp)	RMSR (Fret)	RMSR (Exp)
MODEL						
Null	1061.40(182)	--	--	--	--	--
3 Factor Model	146.37(104)	.86	.873	.894	.080	.068
LX_{Fret3}=$LX_{Expert3}$	219.29(137)	.79	.789	.886	.163	.119

NOTE: **CHI^2** = CHI-square value ; **NFI** = Normed Fit Index; **GFI** = Goodness of Fit Index; **RMSR** = Root Mean Square Residual; **LX**=Lambda X; **PH**=Phi; **3**=3 Factors.

the three- and four-factor model [$\chi^2(93, \underline{N} = 150) = 119.19$, $\underline{p}>.01$, NFI = .90]. Since the four factor model revealed a better fit, invariance was imposed on the factor loadings for the first three factors. This invariance restriction resulted in phi being not positive definite which is an indication of poor fit. The modification indices revealed that the item "examples were presented which were clearly related to the central topic" was problematic (i.e., had high modification indices). This item was excluded from further invariance restrictions in the FOR--expert analyses. The invariance restriction on the first three factors was reassessed revealing gamma congruence [$\chi^2(120, \underline{N} = 150) = 180.74$, $\underline{p}<.01$, NFI = .84; χ^2_{diff} $(27, \underline{N} = 150) = 60.81$, $\underline{p}<.01$]. These models and fit indices are presented in Table 12. Factor loadings for the invariant model are presented in Table 13.

Equality constraints on the factor covariance matrix for the first three factors revealed that beta congruence was absent [$\chi^2(126, \underline{N} = 150) = 198.10$, $\underline{p}<.01$, NFI = .83]. Phi was not positive definite. The factor correlation matrices for the FOR and Expert group are presented in Table 14.

Table 12

Measurement Models Fit Indices For Stacked Runs: Frame of Reference (FOR) and Expert Group

	INDEX					
MODEL	CHI^2 (df)	NFI	GFI (For)	GFI (Exp)	RMSR (For)	RMSR (Exp)
Null	1183.07(132)	--	--	--	--	--
3 Factor Model	150.27(104)	.87	.868	.894	.062	.068
LX For3=LX Expert3	201.70(137)	.83	.799	.887	.103	.088
For 4 Expert 3	119.19(93)	.90	.913	.894	.042	.068
LX For4=LX Expert3*	180.74(120)	.85	.826	.886	.089	.088
PH For4=PH Expert3**	198.10(126)	.83	.825	.862	.115	.138

NOTE: CHI^2 = CHI-square value ; NFI = Normed Fit Index; GFI = Goodness of Fit Index; RMSR = Root Mean Square Residual; LX=Lambda X; PH=Phi; 3=3 Factors; 4=4 Factors.

* Item "examples were presented which were clearly related to the central topic" was not restricted to be invariant.

** Phi is not positive definite

Table 13

Frame of Reference (FOR) and Expert Group Factor Structure Invariance (Maximum Likelihood Estimates)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
Seemed interested in the topic	eye contact tone of voice, facial expression	***.589***	.196	.121	.000
Used examples abstract ideas	how many examples how helpful	-.337	***.857***	.514	.000
Presented the lecture smoothly	transitions between topics smooth	-.321	***.589***	.764	.000
Integrated the material effectively	good transitions summary statement	-.308	***.771***	.409	.000
Followed an outline	include all subtopics, equal time	.173	***1.009***	.258	.000
Followed a logical sequence of thought in his lecture	logical transitions	.000*	***1.000****	.000*	.000*
Was well prepared	number of studies examples, responses to questions	-.269	.519	***.549***	.000
Acted relaxed	body movement, verbal expressions facial expressions	.000*	.000*	***1.000****	.000*
Spoke clearly and distinctly	pronunciation was easy to listen to	.081	.135	***.784***	.000
Spoke with vigor and enthusiam	tone and volume voice, facial expressions	***1.068***	.007	.116	.000
Emphasized important points by raising his voice	tone and volume voice	***.875***	-.039	.210	.000

Table 13

Frame of Reference (FOR) and Expert Group Factor Structure Invariance (Maximum Likelihood Estimates continued)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
Voice was animated	tone of voice	***1.000****	.000*	.000*	.000*
Examples were presented which were clearly related to central topic**	examples useful	-.308(.758)	***.760 (1.085)***	.107(.301)	.000
Used purposeful non-verbal behavior	facial expressions, body movement	***.473***	.275	.444	.000

Fixed for identification
****Invariance not imposed, seperate loadings presented.***

Table 14

Factor Correlation Matrix for Frame of Reference (FOR) Group

	Factor 1	Factor 2	Factor 3	Factor 4
Factor 1	1.00			
Factor 2	.55	1.00		
Factor 3	.85	.55	1.00	
Factor 4	.68	.85	.59	1.00

Factor Correlation Matrix for Expert Group

	Factor 1	Factor 2	Factor 3	Factor 4
Factor 1	1.00			
Factor 2	.33	1.00		
Factor 3	.58	.18	1.00	
Factor 4	.00	.00	.00	1.00

Comparison Group (Comp)

The three-factor model for both the expert and Comparison group fit well as shown in Table 15. The chi-square for the stacked three factor model was $\chi^2(104, N = 157) = 202.20$, $p<.01$, NFI = .87. Equality constraints were then imposed on the factor pattern matrix. This invariance restriction revealed gamma congruence [$\chi^2(137, N = 157) = 247.25$, $p<.001$, NFI = .83; $\chi^2_{diff}(34, N = 157) = 45.05$, $p<.05$]. The factor loadings for the invariant model are presented in Table 16.

Equality constraints were imposed on the factor covariance matrix to assess beta congruence for the comparison group. This invariance restriction revealed that beta congruence was absent [$\chi^2(143, N = 157) = 268.76$, $p<.001$, NFI = .82; $\chi^2_{diff}(6, N = 157) = 21.51$, $p<.01$]. The factor correlation matrices for each of the groups are presented in Table 17.

Alpha Congruence

Alpha congruence was assessed by calculating the traditional accuracy measures of distance and elevation. An Analysis of Variance (ANOVA) on both

Table 15

Measurement Models Fit Indices For Stacked Runs: Comparison (Comp) and Expert Group

		INDEX				
	CHI^2 (df)	NFI	GFI (Comp)	GFI (Exp)	RMSR (Comp)	RMSR (Exp)
MODEL						
Null	1491.93(182)	--	--	--	--	--
3 Factor Model	202.20(104)	.87	.827	.894	.089	.068
LX Comp3=LX Expert3	247.25(137)	.83	.801	.860	.120	.122
PH Comp3=PH Expert3	268.76(143)	.82	.784	.853	.226	.251

NOTE: CHI^2 = CHI-square value ; NFI = Normed Fit Index; GFI = Goodness of Fit Index; RMSR = Root Mean Square Residual; LX=Lambda X; PH=Phi; 3=3 Factors.

Table 16

Comparison (Comp) and Expert Group Factor Structure Invariance(Maximum Likelihood Estimates)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3
Seemed interested in the topic	eye contact tone of voice, facial expression	*.528*	.302	.132
Used examples abstract ideas	how many examples how helpful	.019	*.649*	.147
Presented the lecture smoothly	transitions between topics smooth	-.305	.525	*.811*
Integrated the material effectively	good transitions summary statement	-.123	*.853*	.330
Followed an outline	include all subtopics, equal time	-.006	*.956*	-.039
Followed a logical sequence of thought in his lecture	logical transitions	.000*	*1.000**	.000*
Was well prepared	number of studies examples, responses to questions	.050	*.574*	.453
Acted relaxed	body movement, verbal expressions facial expressions	.000*	.000*	*1.000**
Spoke clearly and distinctly	pronunciation was easy to listen to	.264	.069	*.664*
Spoke with vigor and enthusiasm	tone and volume voice, facial expressions	*1.032*	.035	.111
Emphasized important points by raising his voice	tone and volume voice	*1.246*	.012	-.093

Table 16

Comparison (Comp) and Expert Group Factor Structure Invariance (Maximum Likelihood Estimates continued)

ITEM	FRAME	FACTOR 1	FACTOR 2	FACTOR 3
Voice was animated	tone of voice	***1.000****	.000*	.000*
Examples were presented which were clearly related to central topic	examples useful	.075	***.835***	-.010
Used purposeful non-verbal behavior	facial expressions, body movement	.265	.246	***.497***

* ***Fixed for identification***

Table 17

Factor Correlation Matrix for Comparison (Comp) Group

	Factor 1	Factor 2	Factor 3
Factor 1	1.00		
Factor 2	.78	1.00	
Factor 3	.72	.65	1.00

Factor Correlation Matrix for Expert Group

	Factor 1	Factor 2	Factor 3
Factor 1	1.00		
Factor 2	.19	1.00	
Factor 3	.70	.23	1.00

distance and elevation accuracy, followed by a Scheffe post hoc comparison was performed. Distance accuracy, the sum of the squared differences between subject ratings and true score estimates, was significantly different between each of the trained groups ($\underline{M}_{FRET}$ = .985, $\underline{M}_{FOR}$ = .933) and the comparison group [($\underline{M}_{Comp}$ = 1.145), $\underline{F}(2,233) = 8.15$, $\underline{p}<.001$]. Elevation accuracy, the extent to which a rater approximates true score estimates of the overall performance of ratees, was also significantly different between each of the trained groups ($\underline{M}_{FRET}$ = .597, $\underline{M}_{FOR}$ = .599) and the comparison group [($\underline{M}_{Comp}$ = .843), $\underline{F}(2,233) = 6.54$, $\underline{p}<.01$]. Smaller means reflect greater accuracy.

Distance and elevation accuracy confidence intervals were calculated to further assess alpha congruence. The distance accuracy confidence limits for each of the experimental groups at the 99% confidence interval revealed that zero (0) was not in the confidence intervals [FRET (.2062, 1.764); FOR (.0441, 1.822); Comparison (1.463, 2.145)]. This shows that none of the experimental groups rated the same as the experts. The elevation accuracy confidence limits for each of the experimental groups

at the 99% confidence interval revealed that zero (0) was in the confidence interval for all groups and as such all groups exhibited elevation accuracy [FRET (-.473, 1.667); FOR (-.681, 1.879); Comparison (-.606, 2.29)].

CHAPTER IV

DISCUSSION

The utility of frame of reference training and frame of reference training combined with rater error training was assessed vis-a-vis the newly proposed constructs of gamma and beta congruence. Historically, these training paradigms have been assessed at an alpha level, that is, the correspondence of observed performance ratings with true score estimates provided by expert raters. Alpha congruence however presumes a common expert--trainee metric (beta congruence) and conceptual domain (gamma congruence). Furthermore, this assessment of accuracy does not reveal the intended intermediate effects of the FOR and RET training paradigms. The aim however of achieving beta and gamma congruence is to ultimately achieve alpha congruence.

In the current study, gamma congruence--factor structure congruence with expert raters--was found for both the FOR-trained group and the comparison group. The combined training group, FOR and RET did not

exhibit factor congruence with expert raters. In other words, frames of references were most correctly used, as exhibited by common conceptual domains and measured by expert true score estimates, in the FOR and Comparison group. Although the FRET group performed as well as the FOR on a knowledge test of frames, the FRET group was unable to accurately apply the frames. As previously concluded by Bernardin & Pence (1980) in an assessment of alpha accuracy, RET training is likely to foster a response set that decreases accuracy; in this case accuracy in conceptual domains.

The shared conceptual domain of experts, FOR trainees and the comparison group sheds some new light on the utility of FOR training. The training is designed to "tune in" raters to a common frame of reference. FOR training is based on the assumption that raters have incorrect schema; however this was not found. Findings of the current study show that naive trainees may be already "tuned" to an accurate frame, thus demonstrating that FOR training does not result in more accurate frame utilization than untrained raters. It is cautioned however, that this may be particular to the items and frames used in this study. Perhaps, FOR

training will exhibit greater utility with more cognitively complex frames. It is also important to acknowledge that although classrooms were randomly assigned to experimental conditions, the assumption of random assignment is possibly challenged by background differences among the experimental groups. Although unlikely, training may have had no effect at all and these findings may be a result of systematic differences among psychology classes. Nonetheless, this study shows that in some cases, naive frames, or implicit theories are as accurate as trained frames. Considering such cases, FOR training may not be so necessary as previously thought.

In terms of beta congruence, it was hypothesized that RET training would positively influence scale calibration. The effect of RET on scale calibration could not be examined since members of the FRET group were not evaluating the same constructs as the expert groups. Beta congruence investigation is only meaningful if gamma congruence is present; if gamma congruence is not present, beta congruence is not meaningful. That is, if the conceptual frames of reference are different for experts and trainees, any

quantitative differences in scale calibration are moot and uninterpretable. It was also hypothesized that neither the FOR trainees nor comparison group would exhibit proper scale calibration, since they were not trained to do so. Both groups of trainees did not exhibit beta congruence with the experts: the metric or scale for the groups was different. Beta congruence is measured as differences between latent variable covariances and variances; a close examination of these covariances revealed that both the FOR and the comparison group exhibited much higher correlations between the factors than did the expert group. The different factor covariances signal that trainees are perceiving less difference in the relevant constructs than the experts. It appears that an essential piece missing from training is a method of teaching trainees to attend to and rate aspects of behavior independently with a special emphasis on the importance of trait differentiation. RAT training, proposed in the early eighties, emphasized trait differentiation. More specifically, the training typically consists of a lecture on the multidimensionality of jobs, the need to distinguish among these dimensions, and the need to pay

close attention to performance in light of these dimensions (Bernardin & Pence, 1980). Pulakos's (1984) RAT training additionally gave participants anchors of expected performance at varied scale levels. Given the current study's findings, Pulakos's RAT training should be examined in light of beta and gamma congruence.

Traditional accuracy assessments of elevation and distance accuracy were assessed for didatic purposes despite the absence of gamma and/or beta congruence. Traditional alpha congruence assessments revealed that the trained groups (FOR,FRET) were significantly different from the comparison group. Although not traditionally calculated, confidence intervals of elevation and distance accuracy were computed. The confidence limits revealed that the experimental groups exhibited elevation but not distance accuracy. This follow-up analysis addresses more directly the question of accuracy with the expert group. These findings are troubling on several levels. First, the traditional and confidence limit alpha congruence findings are dissimilar. Second, in light of the above gamma and beta congruence findings the meaningfulness of alpha assessments are challenged. The FRET group did not

share the experts conceptual domain nor scale calibration and the FOR group did not share the later. To date, our training evaluations have been based on alpha-level assessments--which implicitly assume gamma and beta congruence. The lack of such congruence indicates that these assessments are inappropriate. These findings demonstrate that alpha comparisons we make among experts and trainees may be spurious and misleading, especially without gamma and beta congruence information.

The question could be raised concerning the utility of examining beta and gamma congruence when alpha congruence is present. Regardless of alpha congruence, it is important to know whether frames and scale calibrations are shared by raters. As stated earlier, alpha congruence may be a spurious expert--trainee relationship; that is, the processes raters are using differ yet the outcome (i.e.,ratings) appear the same. Furthermore, a close look at the traditional distance and elevation accuracy measures used in the literature reveals that both these indices are relative assessments of accuracy. These measures evaluate differences between ratings of trained groups and

"true" scores provided by experts. In this study, the comparison group was found to be significantly more dissimilar from the experts (i.e., less accurate) than the FOR and FRET groups. However, the trained group's ratings are not the same as the expert group's. In this type of analysis all three groups may be on an absolute level distant from the experts, with one group more distant than the others. Although the trained groups are "less distant", the analyses should not be interpreted as indicating that the trained groups are accurate. The analysis is a relative one. When a non-relative analysis is performed (i.e., confidence intervals) the statistics lead to different conclusions. Thus the question one might raise as to "Why traditional alpha accuracy and not beta and gamma congruence?" is an easy one to answer. Beta and gamma congruence measures do not consider *relative* distance between trained groups, as current alpha assessments do, instead they address more directly the question of correspondence with the expert group. Furthermore, the question of why elevation confidence interval accuracy and not beta and gamma congruence need also be addressed. Elevation accuracy is the extent to which a

rater approximates true score estimates of the overall performance of ratees. Neither gamma and beta congruence estimate overall performance, and as such, the findings here are not discrepant.

Task perception findings are also noteworthy. The comparison group reported greater confidence in their ratings and believed that the lecturer would consider their ratings fair compared to the combined training group. This experimental group however, did not achieve beta or alpha congruence. It is likely that the above task perception difference is a result of the absence of feedback regarding rating error and frames of reference vs. substantial feedback regarding both. The comparison group was not given any indication that their ratings were inaccurate, and thus may have experienced a "blind faith" in their rating ability.

The manipulation check findings are also informative. The experimental groups trained in frame of reference performed less well on a knowledge test of frames than the experts. Although the knowledge test did not include all the trained frames of references the subset of question responses revealed that some of the proper frames were not learned. Those however

trained in rater error performed as well as experts in identifying halo, restriction of range, severity and leniency. It is likely that frames of reference are more difficult to learn than rating errors.

Implications

Both researchers and practitioners can benefit from employing gamma and beta assessments which do not assess relative accuracy but instead focus on actual correspondence. Since these assessments are direct, conclusions regarding congruence are more interpretable and provide greater information than the traditional relative assessments of accuarcy. In addition, beta and gamma congruence measurements more directly assess intermediate training effects, and thus can give us greater insight into the processes that make performance appraisal training effective or ineffective. Furthermore, alpha level comparisons in the absence of beta and gamma assessments are likely to be spurious and misleading. Beta and gamma congruence may also be examined in other person perception situations (e.g., interviewing) to establish whether "judges" of behavior are judging the same constructs on

the same perceptual scale.

From an organizational perspective, these findings along with previous research on alpha accuracy suggest that RET training should not be employed to enhance rater judgments. Furthermore, since FOR training may not result in better frame of reference utilization than is performed by naive raters, its expense and time consuming development may be unwarranted. Training raters who already possess appropriate frames of reference would not be an efficient use of training resources. Organizations may profit from better administrative and development decisions if they focus on behaviors that raters do not naturally exhibit. The current findings suggest that raters should be trained to better distinguish among rating dimensions.

Limitations

When using maximum likelihood estimation in sample sizes less than 100, it frequently happens that estimated unique variance estimates are negative (Hayduk,1987). Heywood cases were encountered in the current investigation. In future assessments of beta and gamma congruence, it would be advantageous to

increase the sample size to lessen the probability of such occurrences. Also as is more typically encountered with smaller sample sizes, the warning "phi is not positive definite" was encountered. This warning establishes that the proposed models' "fit" is poor, often indicating that there are too many factors. Although this was only encountered in the FOR group when equating factor variances, it would be advantageous to reassess the lack of fit with a larger sample size.

Another shortcoming of the study is the use of graduate psychology students as experts. Although generally accepted, much of the research does not report their experience with appraisals. What makes someone an expert? Some of the qualities one would expect experts to possess were evident in this group, others were absent. More specifically, the current study found that experts did report greater experience with having their own performance evaluated, greater familiarity with the job of teaching, were more likely to have heard of FOR and RET before, and had rated or viewed the videotapes previously. The experts did not, however, have more experience rating teachers or

evaluating performance. These findings suggest that stricter criteria for expert status may be warranted. Selection criteria in future research should include rating experience and experience rating the stimulus job in question.

It was necessary to use a contrived laboratory simulation to assess alpha, beta, and gamma congruence. However, a more realistic appraisal setting may produce different results. Typically, a rater evaluates someone he or she has seen before and will see again, ratees are evaluated continually and not from a "snapshot" focused view of behavior. In addition, observations are conducted while performing other, sometimes competing, tasks. The appraisal also typically culminates in some kind of organizational decision, salary increase, transfer, training needed, etc. These natural appraisal processes were not simulated.

Lastly, although classrooms were randomly assigned to experimental conditions, the assumption of random assignment is possibly challenged by background diffferences among the experiemental groups. The comparison group had the fewest working students and

the combined training group had the highest level of education.

Future Research

The constructs of beta and gamma congruence reach beyond the performance appraisal literature and may be applied to other person perception and social judgement literature (e.g., interview decisions). Advancing beyond alpha assessments in examining ratings should provide greater insights into processes that affect judgments, in a variety of person perception literatures.

Congruence assessments also provide researchers with a target standard in which they can examine alpha, beta, and gamma change. Alpha, beta and gamma change may be examined using pre-post training ratings. Pre and post ratings can then be examined vis-a-vis congruence with expert ratings. It would be expected that change is toward greater congruence. An investigation of this nature would provide further insight into training utility.

The current findings suggest that naive raters may already possess accurate frames of reference, but the

inability to differentiate constructs is problematic. RAT training or other training intended to teach raters to differentiate among ratee behaviors should be examined in an alpha, beta, gamma congruence framework. Also, beta congruence should be examined with varied scale formats. The absence of congruence might be remedied by changing scale formats. Perhaps Behaviorally Anchored Rating Scales (Smith & Kendall, 1963) would minimize differential scale perceptions.

FOR training also needs to be examined with more cognitively complex frames than employed in the current study. The constructs of performance utilized were limited to physical aspects of the presenter, aspects of the presented material and organization of content, and general platform skills. More complicated frames may reveal that FOR training has greater utility than the presented findings suggest.

It would also be advantageous to replicate the current study with a larger sample size. The likelihood of Heywood cases decreases as the sample size increases and Heywood cases were evident in the current study.

Chapter V

SUMMARY AND CONCLUSION

The purpose of this study was to introduce the proposed constructs of beta and gamma congruence and to assess promising performance appraisal paradigms, FOR and RET training, in light of these proposed constructs. Gamma congruence was found for both the FOR trained and comparison groups. The combined training group (FRET) did not exhibit factor congruence with expert raters. The above findings demonstrate that FOR training may not result in more accurate frame utilization than is evidenced by untrained raters, and that RET training may actually decrease accuracy in conceptual domains. Beta congruence was absent for the FOR and comparison group. High correlations among the factors signal that trainees are perceiving less differentiation among the relevant constructs than are the experts. Traditional accuracy assessments of elevation and distance accuracy, analogous to alpha level assessments, revealed that the trained groups (FOR, FRET) were significantly different from the comparison group. These findings are explained with

regard to relative rather than absolute accuracy. The importance of examining beta and gamma congruence is also highlighted. Traditional alpha congruence assessments are inappropriate and misleading especially in the absence of beta and gamma congruence. Furthermore, the examination of confidence intervals revealed that the experimental groups exhibited elevation accuracy. The implications of these findings are discussed.

Based on the results of the present study, FOR combined with RET training and FOR training alone are tentatively not recommended. Future research on performance appraisal should further investigate the relationship between performance appraisal training and gamma, beta, and alpha congruence. Confirmation of the present studies findings would suggest the need to redirect performance appraisal training away from frames and toward means of establishing similar scale calibration.

APPENDIX A

Baruch College
Department of Psychology

Summer 1990

The following research study is part of a doctoral dissertation being conducted at Baruch College Department of Psychology by Lynn Gracin, M.A. under the supervision of Roger Millsap, Ph.D.

As a participant you will be asked to rate two videotaped lectures, one on crowding and stress, the other on self fulfilling prophecy. You will also participate in training and be asked to answer some questions about the videotape and some about your background. The results of the study will add to our knowledge of performance appraisal and ratings.

All of your answers will be held in confidence, none of your personal responses will be released to anyone and you may withdraw from the project at any time. The study will take approximately one hour to complete.

I am very appreciative of your help and will be happy to answer any questions you may have. In addition, if you would like, I would be happy to share the findings with you. Thank you for your cooperation.

Sincerely

Lynn Gracin

I have read and understand the information given above and agree to participate in this project.

Signed_____ Date_____

APPENDIX B

LECTURER EVALUATON: SELF - FULFILLING PROPHECY

The presentation that you have just watched consisted of a short lecture and a question-and-answer period. Please indicate how much you agree or disagree with the following statements about the lecture period, question and answer period, and both the lecture and question and answer period.

The following statements refer to the lecture period only.

	Strongly Disagree						Strongly Agree
1. He seemed interested in the topic	1	2	3	4	5	6	7
2. He used clear examples to explain abstract ideas	1	2	3	4	5	6	7
3. He presented the lecture smoothly	1	2	3	4	5	6	7
4. He integrated the material effectively	1	2	3	4	5	6	7
5. He followed an outline	1	2	3	4	5	6	7
6. He followed a logical sequence of thought in his lecture	1	2	3	4	5	6	7

The following statement refer to the question-and-answer period only.

7. He provided relevant answers to the questions	1	2	3	4	5	6	7

The following statements refer to both the lecture and the question and answer period.

8. He was well prepared	1	2	3	4	5	6	7
9. He acted relaxed	1	2	3	4	5	6	7
10. He spoke clearly and distinctly	1	2	3	4	5	6	7
11. He spoke with vigor and enthusiasm	1	2	3	4	5	6	7
12. He emphasized important points by raising his voice	1	2	3	4	5	6	7
13. He looked at the class while speaking	1	2	3	4	5	6	7
14. His voice was animated	1	2	3	4	5	6	7
15. Examples were presented which were clearly related to the central topic	1	2	3	4	5	6	7
16. He used purposeful non-verbal behavior	1	2	3	4	5	6	7

Overall, how would you rate the presentation you just watched?

Very Poor — Excellent

A B C D E F G

(Continue On Next Page)

LECTURER EVALUATON: CROWDING AND STRESS

The presentation that you have just watched consisted of a short lecture and a question-and-answer period. Please indicate how much you agree or disagree with the following statements about the lecture period, question and answer period, and both the lecture and question and answer period.

The following statements refer to the lecture period only.

	Strongly Disagree						Strongly Agree
1. He seemed interested in the topic	1	2	3	4	5	6	7
2. He used clear examples to explain abstract ideas	1	2	3	4	5	6	7
3. He presented the lecture smoothly	1	2	3	4	5	6	7
4. He integrated the material effectively	1	2	3	4	5	6	7
5. He followed an outline	1	2	3	4	5	6	7
6. He followed a logical sequence of thought in his lecture	1	2	3	4	5	6	7

The following statement refer to the question-and-answer period only.

7. He provided relevant answers to the questions	1	2	3	4	5	6	7

The following statements refer to both the lecture and the question and answer period.

8. He was well prepared	1	2	3	4	5	6	7
9. He acted relaxed	1	2	3	4	5	6	7
10. He spoke clearly and distinctly	1	2	3	4	5	6	7
11. He spoke with vigor and enthusiasm	1	2	3	4	5	6	7
12. He emphasized important points by raising his voice	1	2	3	4	5	6	7
13. He looked at the class while speaking	1	2	3	4	5	6	7
14. His voice was animated	1	2	3	4	5	6	7
15. Examples were presented which were clearly related to the central topic	1	2	3	4	5	6	7
16. He used purposeful non-verbal behavior	1	2	3	4	5	6	7

Overall, how would you rate the presentation you just watched?

Very Poor						Excellent	
A	B	C	D	E	F	G	(Stop Here)

APPENDIX C

Please complete the following background information.

1. Which of the following best describes your working situation?

   _____ I am a student
   _____ I am both working and attending school
   _____ I am working and not attending school

2. If you are working, what is your current occupation? (answer only if you work) _____

3. How old are you? _____

4. Are you: _____ Male _____ Female

5. What is the highest educational level you have attained?

   _____ High School
   _____ College Freshmam
   _____ College Sophomore
   _____ College Junior
   _____ College Senior
   _____ B.A./ B.S./B.B.A. or equivalent
   _____ Some Graduate Courses
   _____ M.A./M.S./ M.B.A. or equivalent
   _____ Doctorate

6. How many years of experience have you had in each of the following areas?

   Circle your response to each item:
   0 *None*
   1 *Less than* 1
   2 *More than* 1 *but less than* 3
   3 *More than* 3 *but less than* 5
   4 *More than* 5 *but less than* 10
   5 10 *or more*

a. Evaluating Job Performance _____	0	1	2	3	4	5
b. Evaluating Teachers _____	0	1	2	3	4	5
c. Having your performance evaluated ___	0	1	2	3	4	5

(Continue On Next Page)

7. Have you ever heard of Rater Error Training before? ____ Yes ____ No

8. Have you ever heard of Frame of Reference Training before? ____ Yes ____ No

9. Have you ever rated or viewed this videotape before? ____ Yes ____ No

10. To what extent are the following statements descriptive of your feelings?

Circle your response to each item:
0 *Not at all*
1 *To a very small extent*
2 *To a small extent*
3 *To a moderate extent*
4 *To a great extent*
5 *To a very great extent*

a. I feel confident that my ratings are accurate ___	0	1	2	3	4	5
b. The lecturer would consider my ratings fair ___	0	1	2	3	4	5
c. I feel comfortable rating teachers _____	0	1	2	3	4	5
d. I am familiar with the job of teaching ___	0	1	2	3	4	5
e. I thought the training was interesting ___	0	1	2	3	4	5
f. The training helped me make a more accurate evaluation _____	0	1	2	3	4	5
g. I anticipate that the training will help me make accurate evaluations in the future ___	0	1	2	3	4	5

(Continue On Next Page)

Expert Knowledge of Appraisal Form

The following questions refer to both performance appraisal errors and the information you have just read. Please place a check in the box next to the best answer.

1) Who is committing halo error rater 1 or 2?

Rater							
❑ 1	7	7	6	7	6	7	6
❑ 2	7	4	1	2	3	1	6

2) Who is committing restriction of range rater 1 or 2?

Rater							
❑ 1	4	5	4	4	6	5	5
❑ 2	7	5	2	3	4	6	7

3) Who is probably giving too severe ratings rater 1 or 2 (A rating of 1 is low and 7 is high)?

Rater							
❑ 1	1	2	1	3	1	1	1
❑ 2	1	3	5	7	2	3	1

4) Who is probably giving too lenient ratings rater 1 or 2 (A rating of 1 is low and 7 is high)?

Rater							
❑ 1	7	7	6	7	6	7	6
❑ 2	7	6	5	3	6	7	4

5) True or False: to evaluate "he seemed interested in the topic" you should look at:

a. eye contact. . . . . ❑True ❑False

b. tone of voice. . . . . ❑True ❑False

c. facial expression. . . . . ❑True ❑False

d. whether he provided a summary statement. . . . . ❑True ❑False

(Continue On Next Page)

6) **True or False: to evaluate "he used clear examples to explain abstract ideas" you should look at:**

a. whether the lecture was smooth. . . . . ❑True ❑False

b. how many real life examples. . . . . ❑True ❑False

c. how helpful examples were in getting points across. . ❑True ❑False

7) **True or False: to evaluate "he followed a logical sequence of thought in his lecture" you should look at:**

a. whether movement from 1 idea to the next seemed logical. . . . . ❑True ❑False

b. whether answers were clear. . . . . ❑True ❑False

8) **True or false: to evaluate "he was well prepared" you should look at:**

a the number of research studies cited. . . . . ❑True ❑False

b. number of examples used in the lecture. . . . . ❑True ❑False

c. responses to questions. . . . . ❑True ❑False

d. facial expressions. . . . . ❑True ❑False

9) **True of False: to evaluate "he acted relaxed" you should look at:**

a. body movement. . . . . ❑True ❑False

b. verbal presentation. . . . . ❑True ❑False

Frame of Reference and Rater Error Training Group

The following questions refer to the training you just received.. Please place a check in the box next to the best answer.

1) **Who is committing halo error rater 1 or 2?**

Rater							
❑ *1*	*7*	*7*	*6*	*7*	*6*	*7*	*6*
❑ *2*	*7*	*4*	*1*	*2*	*3*	*1*	*6*

2) **Who is committing restriction of range rater 1 or 2?**

Rater							
❑ *1*	*4*	*5*	*4*	*4*	*6*	*5*	*5*
❑ *2*	*7*	*5*	*2*	*3*	*4*	*6*	*7*

3) **Who is probably giving too severe ratings rater 1 or 2 (A rating of 1 is low and 7 is high)?**

Rater							
❑ *1*	*1*	*2*	*1*	*3*	*1*	*1*	*1*
❑ *2*	*1*	*3*	*5*	*7*	*2*	*3*	*1*

4) **Who is probably giving too lenient ratings rater 1 or 2 (A rating of 1 is low and 7 is high)?**

Rater							
❑ *1*	*7*	*7*	*6*	*7*	*6*	*7*	*6*
❑ *2*	*7*	*6*	*5*	*3*	*6*	*7*	*4*

5) **True or False: to evaluate "he seemed interested in the topic" you should look at:**

a. eye contact. . . . . ❑True ❑False

b. tone of voice. . . . . ❑True ❑False

c. facial expression. . . . . ❑True ❑False

d. whether he provided a summary statement. . . . . ❑True ❑False

(Continue On Next Page)

6) **True or False: to evaluate "he used clear examples to explain abstract ideas" you should look at:**

a. whether the lecture was smooth. . . . . ❑True ❑False

b. how many real life examples. . . . . ❑True ❑False

c. how helpful examples were in getting points across. . ❑True ❑False

7) **True or False: to evaluate "he followed a logical sequence of thought in his lecture" you should look at:**

a. whether movement from 1 idea to the next seemed logical. . . . . ❑True ❑False

b. whether answers were clear. . . . . ❑True ❑False

8) **True or false: to evaluate "he was well prepared" you should look at:**

a the number of research studies cited. . . . . ❑True ❑False

b. number of examples used in the lecture. . . . . ❑True ❑False

c. responses to questions. . . . . ❑True ❑False

d. facial expressions. . . . . ❑True ❑False

9) **True of False: to evaluate "he acted relaxed" you should look at:**

a. body movement. . . . . ❑True ❑False

b. verbal presentation. . . . . ❑True ❑False

Frame of Reference Training Group

The following questions refer to the training you just received. Please place a check in the box next to the best answer.

1) **True or False: to evaluate "he seemed interested in the topic" you should look at:**

a. eye contact. . . . . ❑True ❑False

b. tone of voice. . . . . ❑True ❑False

c. facial expression. . . . . ❑True ❑False

d. whether he provided a summary statement. . . . . ❑True ❑False

2) **True or False: to evaluate "he used clear examples to explain abstract ideas" you should look at:**

a. whether the lecture was smooth. . . . . ❑True ❑False

b. how many real life examples. . . . . ❑True ❑False

c. how helpful examples were in getting points across. . ❑True ❑False

3) **True or False: to evaluate "he followed a logical sequence of thought in his lecture" you should look at:**

a. whether movement from 1 idea to the next seemed logical. . . . . ❑True ❑False

b. whether answers were clear. . . . . ❑True ❑False

4) **True or false: to evaluate "he was well prepared" you should look at:**

a the number of research studies cited. . . . . ❑True ❑False

b. number of examples used in the lecture. . . . . ❑True ❑False

c. responses to questions. . . . . ❑True ❑False

d. facial expressions. . . . . ❑True ❑False

(Continue On Next Page)

5) **True of False: to evaluate "he acted relaxed" you should look at:**

a. body movement. . . . . ❑True ❑False

b. verbal presentation. . . . . ❑True ❑False

APPENDIX D

Expert Group Instructions

I am asking you to view two videotapes, complete two lecture evaluations, some background information and a "knowledge of appraisal form".

The lecturer you are about to observe , the first videotape, was asked to lecture on the topic of crowding and stress and to include three subtopics in his lecture: cultural differences, sex differences, and personal space.

Before viewing the first lecture , please review the items and scales in the evaluation form on the next page. Note that the evaluation form has four sections. Section 1 refers to the lecture period only, section 2 refers to the question and answer period only, and section 3 and 4 refer to both the lecture and question-and-answer period.

(View videotape on crowding and stress two times. **Replay videotape as many times as needed.** Rate videotape).

Do not continue until you have completed rating the crowding and stress lecture.

Once you read on , do not change your original ratings.

As you read on please do not change your original ratings.

The following outlines what other "experts" looked at when they evaluated the lecture. Please go through the evaluation form statement by statement and pay attention to the specific conditions looked at by other expert raters.

Statement #1: He seemed interested in the topic. Rating 5

The experts looked :

1) eye contact
2) tone of voice
3) facial expression

eye contact (+)

Although he read the lecture he kept looking up at the audience to personally convey the message to them.

tone of voice (-)

A criticism by the experts was his monotone voice. There was not much excitement in his voice.

facial expression (-)

Another criticism was that, if you noticed, he never smiled.

Based on these three points, the experts gave him a 5.

Statement #2: He used clear examples to explain abstract ideas. Rating 6

The experts looked :

1)how many real life examples

2)how helpful they were in getting points across

how many (+)

He offered quite a few examples such as: how we react to crowding on buses, how we react to crowding at cocktail parties, how we react to crowding at football games, the effects of small rooms on jury trials.

how useful (+)

The experts thought they were good but not good enough to offer a strongly agree.

Based on these points the experts gave him a 6.

Statement #3: He presented the lecture smoothly. Rating 6

The experts looked :

1)transitions from 1 topic to the next. Were the transitions smooth?

smooth transitions(+)

The experts rated the lecturer fairly high on this dimension.

Examples of smooth transitions are:

1) Transition from discussion of animal studies to human research.

2) Transition from general human research to sex differences in human research.

Based on these points the experts gave him a 6.

Statement #4: He integrated the material effectively. Rating 5

The experts looked :

1)good transitions

2)did he provide a summary statement that tied together the subtopics.

good transitions (+) Already rated 6.

summary statement (+/-)

At the end of the lecture he did provide a summary statement. He mentioned: animal research, human research, personal space. But, the summary was short and not really clear.

Based on these points the experts gave him a 5.

Statement #5: He followed an outline. Rating 3.

Note: He was asked to include cultural differences, sex differences, and personal space.

The experts looked :

1)did he include all subtopics.

2)did he devote the same amount of time to each.

all 3 topics (-)

No, he lectured on sex differences, personal space, but not cultural differences (just included one sentence about rural vs. Tokyo-Houston)

equal time (-)

No, only during Q&A.

Based on these points the experts gave him a 3.

Statement #6: He followed a logical sequence of thought in his lecture. Rating 6

The experts looked :

1)whether movement from 1 idea to the next seemed logical.

logical transitions (+)

The experts rated the lecturer fairly high on this dimension.

Examples of logical transition are:

1)transition from general human research to sex differences in human research.

2)transition from discussion of animal studies to human research.

Based on these points the experts gave him a 6.

Statement #7: He provided relevant answers to the questions. Rating 3.

The experts looked :

1)whether or not the answers were:

a.clear

b.sufficient

c.convincing

The experts rated the lecturer rather low on this dimension. The last question is used as an example of why--"Is crowding always stressful?" Answer-Probably not. It depends on actual density, perceived levels of crowding, and frequency of violations of personal space.

clear (-)

The response is concise but not clear.

sufficient(-)

It is too short. He should have elaborated more and explained what he meant by each of the three terms and how they related to each other.

Question#2 also is used as an example: What accounts for cultural differences reaction to stress?

Answer-In different cultures, different levels of crowding are perceived as stressful. It has to do with adapting to your environment. If you spend time in a crowded

place it no longer seems crowded to you. Most cultures probably adapt to their own levels of crowding.

convincing (-)

The answer indicates that he is not sure and has not read enough on the topic to support his statement.

Based on his failure to be clear, sufficient and convincing on most of his answers the experts gave him a 3 on this dimension.

Statement #8: He was well prepared. Rating 6.

The experts looked :

1)the number of research studies cited.

2)number of examples used in the lecture.

3)responses to questions.

research studies and examples (+)

He used this type of information quite often. For example:

1) Calhoun's study with rats (1962), 2) research on jury trials, 3)over-crowding on buses, at football games, and cocktail parties.

responses to questions (-)

While the experts rated him very favorably on the above, his rating here suffered from his inadequate responses to questions.

Based on these points the experts gave him a 6.

Statement #9: He acted relaxed. Rating 3.

The experts looked :

1)body movement

2)verbal presentation

3)facial expression

Body Movements (+)

There were no apparent displays of being nervous. He did not tremble or play nervously with his hands.

Verbal Presentation(-)

The way he read the lecture indicated that he was not comfortable to ad lib and be spontaneous.

Facial Expression(-)

He never smiled once which would have indicated he was relaxed. The experts gave him a 3.

Statement #10: He spoke clearly and distinctly. Rating 6.

The experts looked :

1)pronunciation of words and phrases. Was he easy to listen to.

Pronunciation (+)

His words and phrases were clear and easy to understand.

Easy to listen to (-)

He spoke somewhat quietly and did not really project his statements. The experts gave him a 6.

Statement #11: He spoke with vigor and enthusiasm. Rating 4.

The experts looked :

1)facial expression

2)tone and volume of voice.

Facial Expression(+)

While he made an effort to look at the audience, he conveyed little excitement through facial expression.

Tone and volume of voice(-)

While the tone and volume of his voice were not powerful, he did manage to get his message across.

The experts gave him a 4.

Statement #12: He emphasized important points by raising his voice. Rating 4.

The experts looked :

1)tone of voice.

2)volume of voice.

Tone and volume of voice (-)

On occasion he used this technique, however, you had to listen closely to observe it.

Examples: 1)he raised the tone of his voice to direct attention to: a new topic he was discussing; 2)he did this to let you know he was going to talk about human research and not rat research, 3)he did this to direct your attention to a new term-personal space.

The experts gave him a 4.

Statement #13: He looked at the class while speaking. Rating 6.

Look at :

1)eye contact.

Eye Contact (+)

Although he read the lecture, he kept looking up at the audience to personally convey the message to them.

Statement #14: His voice was animated. Rating 3.

Look at :

1)tone of voice.

Tone of voice (-)

He had a monotone voice and showed little excitement.

Statement #15: Examples were presented which were clearly related to the topic. Rating 6.

Look at :

1)whether the examples were useful.

Examples (+)

The examples were good but not good enough to offer a strongly agree.

Statement #16: He used purposeful non-verbal behavior. Rating 4.

Look at :

1)facial expression

2)body movement.

Facial Expression(-)

He conveyed little through facial expression. He never smiled.

Body Movement(-)

He stayed close to the podium and did not purposefully use his hands or body to stress points. During the question and answer period his body movements were more animated and involving.

You are now going to view and rate a lecture on self-fulfilling prophecy. Please apply the appropriate specific conditions looked at by other expert raters. The lecturer was asked to include teacher expectations and grades, parents expectations and sex roles, and first impressions in developing friendships.

(View videotape on self-fulfilling prophecy two times. Replay videotape as many times as needed. Rate videotape).

APPENDIX E

FRAME OF REFERENCE TRAINING

I am asking you to fill out this lecture evaluation and the attached background information after you have observed the videotaped lecture. Some people when they fill out such questionnaires, tend to respond in certain ways which could make their evaluations inaccurate. Today, we are going to spend some time talking about what "experts" look at when they evaluate teacher performance. First, lets look at the rating scale (highlight scale anchors, values, and dimensions).

Now, pay close attention to the following lecturer's performance and evaluate it using the form in front of you. The lecturer was asked to lecture on the topic of crowding and stress and to include three subtopics in his lecture: cultural differences, sex differences, and personal space. We will discuss how experts evaluated the performance.

(Show videotape on crowding and stress. Have students rate videotape). Please do not erase your original ratings.

We will go through the evaluation form statement by statement and discuss the specific conditions looked at by the expert raters.

Statement #1: He seemed interested in the topic.

Rating 5

The experts looked : 1)eye contact
2)tone of voice
3)facial expression

eye contact (+) Although he read the lecture he kept looking up at the audience to personally convey the message to them.

tone of voice (-) A criticism by the experts was his monotone voice. There was not much excitement in his voice.

facial expression (-) Another criticism was that, if you noticed, he never smiled.

Based on these three points, the experts gave him a 5.

Statement #2: He used clear examples to explain abstract ideas.

Rating 6

The experts looked : 1)how many real life examples
2)how helpful they were in getting points across

how many (+) He offered quite a few examples such as:
How we react to crowding on buses.
How we react to crowding at cocktail parties.
How we react to crowding at football games.
The effects of small rooms on jury trials.

how useful (+) The experts thought they were good but not good enough to offer a strongly agree.

Based on these points the experts gave him a 6.

Statement #3: He presented the lecture smoothly. Rating 6

The experts looked : 1)transitions from 1 topic to the next. Were the transitions smooth?

smooth transitions(+) The experts rated the lecturer fairly high on this dimension. Examples of smooth

transitions are:

1) Transition from discussion of animal studies to human research.
2) Transition from general human research to sex differences in human research.

Based on these points the experts gave him a 6.

Statement #4: He integrated the material effectively. Rating 5

The experts looked : 1)good transitions
2)did he provide a summary statement that tied together the subtopics.

good transitions (+) Already rated 6.

summary statement (+/-) At the end of the lecture he did provide a summary statement. He mentioned: animal research, human research, personal space. But, the summary was short and not really clear.

Based on these points the experts gave him a 5.

Statement #5: He followed an outline. Rating 3.

Note: He was asked to include cultural differences, sex differences, and personal space.

The experts looked : 1)did he include all sub topics.

2)did he devote the same amount of time to each.

all 3 topics (-) No, he lectured on sex differences, personal space, but not cultural differences (just included one sentence about rural vs. Tokyo-Houston)

equal time (-) No, only during Q&A.

Based on these points the experts gave him a 3.

Statement #6: He followed a logical sequence of thought in his lecture. Rating 6

The experts looked :

1)whether movement from 1 idea to the next seemed logical.

logical transitions

(+) The experts rated the lecturer fairly high on

this dimension. Examples of logical transition are:

1) transition from general human research to sex differences in human research.
2) transition from discussion of animal studies to human research.

Based on these points the experts gave him a 6.

Statement #7: He provided relevant answers to the questions.

Rating 3.

The experts looked :

1) whether or not the answers were:
   a. clear
   b. sufficient
   c. convincing

The experts rated the lecturer rather low on this dimension. The last question is used as an example of why--"Is crowding always stressful?" Answer-Probably not. It depends on actual density, perceived levels of crowding, and frequency of violations of personal space.

clear (-) The response is concise but not clear.

sufficient(-) It is too short. He should have elaborated more and explained what he meant by each of the three terms and how they related to each other.

Question#2 What accounts for cultural differences in reaction to stress?

Answer-In different cultures, different levels of crowding are perceived as stressful. It has to do with adapting to your environment. If you spend time in a crowded place it no longer seems crowded to you. Most cultures probably adapt to their own levels of crowding.

convincing (-) The answer indicates that he is not sure and has not read enough on the topic to support his statement.

Based on his failure to be clear, sufficient and convincing on most of his answers the experts gave him a 3 on this dimension.

Statement #8: He was well prepared. Rating 6.

The experts looked : 1) the number of research studies cited.
2) number of examples used in the lecture.

3)responses to questions.

research studies and examples (+) He used this type of information quite often. For example: 1) Calhoun's study with rats (1962), 2) research on jury trials, 3)over-crowding on buses, at football games, and cocktail parties.

responses to questions (-) While the experts rated him very favorably on the above, his rating here suffered from his inadequate responses to questions.

Based on these points the experts gave him a 6.

Statement #9: He acted relaxed. Rating 3.

The experts looked : 1)body movement
2)verbal presentation
3)facial expression

Body Movements (+) There were no apparent displays of being

nervous. He did not tremble or play nervously with his hands.

Verbal Presentation

(-)The way he read the lecture indicated that he was not comfortable to ad lib and be spontaneous.

Facial Expression

(-)He never smiled once which would have indicated he was relaxed.

The experts gave him a 3.

Statement #10: He spoke clearly and distinctly. Rating 6.

The experts looked : 1)pronunciation of words and phrases. Was he easy to listen to.

Pronunciation (+) His words and phrases were clear and easy to understand.

(-) He spoke somewhat quietly and did not really project his statements.

The experts gave him a 6.

Statement #11: He spoke with vigor and enthusiasm. Rating 4.

The experts looked : 1)facial expression
2)tone and volume of voice.

Facial Expression

(+) while he made an effort to look at the audience, he conveyed little excitement through facial expression.

Tone and volume of voice

(-) while the tone and volume of his voice were not powerful, he did manage to get his message across.

The experts gave him a 4.

Statement #12: He emphasized important points by raising his voice. Rating 4.

The experts looked : 1)tone of voice.

2)volume of voice.

Tone and volume

of voice (-) on occasion he used this technique, however, you had to listen closely to observe it. Examples: 1)he raised the tone of his voice to direct attention to: a new topic he was discussing; 2)he did this to let you know he was going to talk about human research and not rat research, 3)he did this to direct your attention to a new term-personal space.

The experts gave him a 4.

*Statement #13: He looked at the class while speaking.

Rating 6.

Look at : 1)eye contact.

Eye Contact (+) Although he read the lecture, he kept looking up at the audience to personally convey the message to them.

*Statement #14: His voice was animated. Rating 3.

Look at : 1)tone of voice.

Tone of voice (-) He had a monotone voice and showed little excitement.

*Statement #15: Examples were presented which were clearly related to the topic. Rating 6.

Look at : 1)whether the examples were useful.

Examples (+) The examples were good but not good enough to offer a strongly agree.

*Statement #16: He used purposeful non-verbal behavior. Rating 4.

Look at : 1)facial expression
2)body movement.

Facial Expression

(-) He conveyed little through facial expression. He never smiled.

Body Movement

(-) He stayed close to the podium and did not purposefully use his hands or body to stress

points. During the question and answer period his body movements were more animated and involving.

*Statements 13-16 are from Murphy's behavioral scale, true scores are from Murphy and frames of reference are from the author.

Any questions? We are now going to view and rate a lecture on self-fulfilling prophecy. The lecturer was asked to include teacher expectations and grades, parent's expectations and sex roles, first impressions in developing friendships.

APPENDIX F

Frame of Reference/Rater Error Training

I am asking you to fill out this lecture evaluation and the attached background information after you have observed the videotaped lecture. Some people when they fill out such questionnaires, tend to respond in certain ways which could make their evaluations inaccurate. Today, we are going to spend some time talking about what "experts" look at when they evaluate teacher performance and errors in evaluation. First, lets look at the rating scale (highlight scale anchors, values, and dimensions).

Now pay close attention to the following lecturer's performance and evaluate it using the form in front of you. The lecturer was asked to lecture on the topic of crowding and stress and to include three subtopics in his lecture: cultural differences, sex differences, and personal space. We will discuss both evaluation errors and how experts evaluated the performance.

(Show videotape on crowding and stress. Have students rate videotape). Please do not erase your original ratings.

First we are going to spend some time talking about errors to avoid when evaluating teacher performance.

The most common mistake people make occurs when they fail to distinguish between different aspects of the lecture. Just like everything else we have opinions about--movies, clothing, books, records, wine--there are a number of characteristics on which we can judge a lecture.

For example, let's consider the situation in which we are evaluating a movie. Let's consider the movie "Teenage Mutant Ninja Turtles". If someone were to ask us if we liked the movie we might answer by saying something like this:

"The plot was silly".

"I sort of liked Michaelangelo's performance".

"The theme was not realistic at all".

"The background music was excellent".

The point is that we judge the movie on different characteristics **which we have observed.** On certain characteristics we evaluated the movie as very good (the music); on certain characteristics we evaluated the movie as poor (the theme); and on certain characteristics such as Michaelangelo's performance we evaluated

the movie as moderately good. These judgments which we have made are based on what we personally observed.

Observations are most accurate when :

1) Like above, we distinguish among different aspects of the movie (acting, plot, music, etc.). When we do not distinguish among the various aspects we are committing halo error.

Who is committing halo error rater 1 or 2?

Rater							
1	1	2	1	1	1	2	2
2	7	4	1	3	4	1	6

2) Like the movie ratings, we avoid giving all ratings that fall around one part of the scale or in the middle of the scale, unless that is what is deserved. In the movie example, many would disagree with my ratings if I said "The movie was average", "The acting was O.K.", "The plot was O.K.". In other words, when

we rate, we should use the whole rating scale and base our ratings on what we observed. In addition, don't be afraid to use extreme ratings like "the plot was very good".

When we do not use the entire scale we are restricting the range of ratings?

Who is committing restriction of range rater 1 or 2?

Rater							
1	3	2	3	2	2	3	3
2	7	6	4	5	3	2	7

3) We don't give ratings that are more favorable or more severe than what is deserved.

Who is probably giving too severe ratings rater 1 or 2 (A rating of 1 is low and 7 is high)?

Rater							
1	1	2	1	5	1	1	1
2	7	6	4	5	6	1	2

Who is probably giving too lenient ratings rater 1 or 2 (A rating of 1 is low and 7 is high)?

Rater							
1	7	7	7	6	6	7	7
2	7	6	4	5	2	4	6

Take a minute to look at your ratings. Did you commit any of the errors we discussed?

Now, we will go through the evaluation form statement by statement and discuss the specific conditions looked at by the expert raters.

Statement #1: He seemed interested in the topic.

Rating 5

The experts looked : 1)eye contact
2)tone of voice
3)facial expression

eye contact (+) Although he read the lecture he kept looking up at the audience to personally convey the message to them.

tone of voice (-) A criticism by the experts was his monotone voice. There was not much excitement in his voice.

facial expression (-) Another criticism was that, if you noticed, he never smiled.

Based on these three points, the experts gave him a 5.

Statement #2: He used clear examples to explain abstract ideas.

Rating 6

The experts looked : 1)how many real life examples
2)how helpful they were in getting points across

how many (+) He offered quite a few examples such as:
How we react to crowding on buses.
How we react to crowding at cocktail parties.
How we react to crowding at football games.
The effects of small rooms on jury trials.

how useful (+) The experts thought they were good but not good enough to offer a strongly agree.

Based on these points the experts gave him a 6.

Statement #3: He presented the lecture smoothly. Rating 6

The experts looked : 1) transitions from 1 topic to the next. Were the transitions smooth?

smooth transitions(+) The experts rated the lecturer fairly high on this dimension. Examples of smooth transitions are:

1) Transition from discussion of animal studies to human research.
2) Transition from general human research to sex differences in human research.

Based on these points the experts gave him a 6.

Statement #4: He integrated the material effectively. Rating 5

The experts looked : 1)good transitions
2)did he provide a summary statement that tied together the subtopics.

good transitions (+) Already rated 6.

summary statement (+/-) At the end of the lecture he did provide a summary statement. He mentioned: animal research, human research, personal space. But, the summary was short and not really clear.

Based on these points the experts gave him a 5.

Statement #5: He followed an outline. Rating 3.

Note: He was asked to include cultural differences, sex differences, and personal space.

The experts looked : 1)did he include all sub topics.
2)did he devote the same amount of time to each.

all 3 topics (-) No, he lectured on sex differences, personal space, but not cultural differences (just included one sentence about rural vs. Tokyo-Houston)

equal time (-) No, only during Q&A.

Based on these points the experts gave him a 3.

Statement #6: He followed a logical sequence of thought in his lecture. Rating 6

The experts looked :

1) whether movement from 1 idea to the next seemed logical.

logical transitions

(+) The experts rated the lecturer fairly high on this dimension. Examples of logical transition are:

1) transition from general human research to sex differences in human research.

2) transition from discussion of animal studies to human research.

Based on these points the experts gave him a 6.

Statement #7: He provided relevant answers to the questions.

Rating 3.

The experts looked :

1) whether or not the answers were:

a. clear

b. sufficient

c. convincing

The experts rated the lecturer rather low on this dimension. The last question is used as an example of why--"Is crowding always stressful?" Answer-Probably not. It depends on actual density, perceived levels of crowding, and frequency of violations of personal space.

clear (-) The response is concise but not clear.

sufficient(-) It is too short. He should have elaborated more and explained what he meant by each of the three terms and how they related to each other.

Question#2 What accounts for cultural differences in reaction to stress?

Answer-In different cultures, different levels of crowding are perceived as stressful. It has to do with adapting to your

environment. If you spend time in a crowded place it no longer seems crowded to you. Most cultures probably adapt to their own levels of crowding.

convincing (-) The answer indicates that he is not sure and has not read enough on the topic to support his statement.

Based on his failure to be clear, sufficient and convincing on most of his answers the experts gave him a 3 on this dimension.

Statement #8: He was well prepared. Rating 6.

The experts looked : 1) the number of research studies cited.
2) number of examples used in the lecture.
3) responses to questions.

research studies and examples (+) He used this type of information quite often. For example: 1) Calhoun's study with rats (1962), 2) research on jury trials, 3) over-crowding on buses, at football games, and cocktail parties.

responses to questions (-) While the experts rated him very favorably on the above, his rating here suffered from his inadequate responses to questions.

Based on these points the experts gave him a 6.

Statement #9: He acted relaxed. Rating 3.

The experts looked : 1) body movement
2) verbal presentation
3) facial expression

Body Movements (+) There were no apparent displays of being nervous. He did not tremble or play nervously with his hands.

Verbal Presentation

(-) The way he read the lecture indicated that he was not comfortable to ad lib and be spontaneous.

Facial Expression

(-)He never smiled once which would have indicated he was relaxed.

The experts gave him a 3.

Statement #10: He spoke clearly and distinctly. Rating 6.

The experts looked : 1)pronunciation of words and phrases. Was he easy to listen to.

Pronunciation (+) His words and phrases were clear and easy to understand.

(-) He spoke somewhat quietly and did not really project his statements.

The experts gave him a 6.

Statement #11: He spoke with vigor and enthusiasm. Rating 4.

The experts looked : 1)facial expression
2)tone and volume of voice.

Facial Expression

(+) while he made an effort to look at the audience, he conveyed little excitement through facial expression.

Tone and volume of voice (-) while the tone and volume of his voice were not powerful, he did manage to get his message across.

The experts gave him a 4.

Statement #12: He emphasized important points by raising his voice. Rating 4.

The experts looked : 1)tone of voice.
2)volume of voice.

Tone and volume of voice (-) on occasion he used this technique, however, you had to listen closely to observe it. Examples: 1)he raised the tone of his voice to direct attention to: a new topic he was discussing; 2)he did this to let you know he was going to talk about human research and not

rat research, 3)he did this to direct your attention to a new term-personal space.

The experts gave him a 4.

*Statement #13: He looked at the class while speaking. Rating 6.

Look at : 1)eye contact.

Eye Contact (+) Although he read the lecture, he kept looking up at the audience to personally convey the message to them.

*Statement #14: His voice was animated. Rating 3.

Look at : 1)tone of voice.

Tone of voice (-) He had a monotone voice and showed little excitement.

*Statement #15: Examples were presented which were clearly related to the topic. Rating 6.

Look at : 1)whether the examples were useful.

Examples (+) The examples were good but not good enough to offer a strongly agree.

*Statement #16: He used purposeful non-verbal behavior.

Rating 4.

Look at : 1) facial expression

2) body movement.

Facial Expression

(-) He conveyed little through facial expression. He never smiled.

Body Movement

(-) He stayed close to the podium and did not purposefully use his hands or body to stress points. During the question and answer period his body movements were more animated and involving.

*Statements 13-16 are from Murphy's behavioral scale, true scores are from Murphy and frames of reference are from the author.

Any questions? We are now going to view and rate a lecture on self-fulfilling prophecy. The lecturer was asked to include teacher expectations and grades, parents' expectations and sex roles, first impressions in developing friendships.

APPENDIX G

CROWDING AND STRESS LECTURE

Why is crowding a topic of interest?

Crowding constitutes an aspect of one of the major social issues of our time. Consider this:
The population of the world increases by some 100 million people every year, and by 10 billion every 10 years. Add to this the point that 50 percent of the United States population lives on just one percent of the land, and you can see why crowding has become a problem of increasing proportions.

Why is crowding such an unpleasant experience?

Three recent explanations are offered:

1) STIMULUS OVERLOAD EXPLANATION

People may find that high density situations produce so much stimulation that they are unable to process the information effectively. This inability to sort out and consider competing stimuli leads to psychological stress.

2) THE LOSS OF CONTROL EXPLANATION

When there are many individuals in a small area, people may feel they have less control over the situation and that their behavioral freedom has been reduced.

3) THE FOCUS OF ATTENTION EXPANATION

When we are with many people in a small area, others may approach us closely--invading our personal space. When this happens a typical reaction is to become psychologically and physiologically aroused. According to this model, people will become motivated to determine why they are aroused. If the arousal is attributed to the presence of others--"That guy is standing really close to me"--they will experience crowding. If however, their attention is focused on some other aspect--"I drank too much beer", "The wallpaper is too busy and driving me nuts", "I ate too much"-crowding will not be experienced.

Consequences of Crowding

There are consequences of crowding beyond a negative emotional response. The effects of crowding are surprisingly powerful. Higher rates of illness are related to crowding in a variety of settings. For example, it has been found that students who live in dormitories of higher density pay more visits to the student health center than those living in low density housing. The death rates of inmates in a state psychiatric prison were also found to be related to density--as the population grew so did the per-capita death rate; and as the population fell the death rate

followed.

Crowding also effects interpersonal attraction: People are less likely to like one another under conditions of high density than when the density is lower. For example, people with three roomates liked the roomates less than people with two roomates.

Crowding also leads to a decrease in helping behavior, an increase in aggression, and declines performance levels on various tasks.

But, living in New York is not so dim..it is though that people eventually adopt to crowded conditions, particularly when their neighbors are living under the same circumstances.

Any questions? We are now going to view and rate a lecture on self-fulfilling prophecy. The lecturer was asked to include teacher expectations and grades, parents expectations and sex roles, first impressions in developing friendships.

APPENDIX H

Expert Group Covariance Matrix

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12	Var13	Var14
Var1	1.118													
Var2	.557	1.649												
Var3	.531	.979	1.761											
Var4	.459	1.035	.969	1.772										
Var5	.492	1.126	.765	1.003	2.607									
Var6	.411	1.094	.953	1.100	4.459	1.910								
Var7	.261	.716	.670	.896	.602	.879	1.885							
Var8	.338	.555	1.068	.620	.105	.349	.713	1.973						
Var9	.653	.573	.962	.524	.335	.391	.669	1.003	2.042					
Var10	.595	.394	.583	.433	.391	.477	.402	.793	.707	1.344				
Var11	.461	.207	.446	.221	.304	.311	.502	.732	.602	.895	1.280			
Var12	.497	.197	.364	.220	.313	.301	.125	.553	.552	.848	.702	1.049		
Var13	.229	.864	.537	.789	.937	.969	.555	.169	.273	.082	.022	.102	1.212	
Var14	.466	.618	.527	.530	.431	.515	.737	.953	.951	.811	.607	.713	.524	1.800

KEY

Var1	Seemed interested in the topic
Var2	Used examples abstract ideas
Var3	Presented the lecture smoothly
Var4	Integrated the material effectively
Var5	Followed an outline
Var6	Followed a logical sequence of thought in his lecture
Var7	Was well prepared
Var8	Acted relaxed
Var9	Spoke clearly and distinctly
Var10	Spoke with vigor and enthusiasm
Var11	Emphasized important points by raising his voice
Var12	Voice was animated
Var13	Examples were presented which were clearly related to the central topic
Var14	Used purposeful non-verbal behavior

Frame of Reference and Rater Error Training Group Covariance Matrix

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12	Var13	Var14
Var1	1.394													
Var2	.754	1.389												
Var3	.686	.681	1.378											
Var4	.695	.792	.872	1.641										
Var5	.589	.283	.677	.915	1.794									
Var6	.640	.659	.726	.919	.717	1.467								
Var7	.956	.896	.766	.840	.572	.730	1.756							
Var8	.830	.483	.874	.599	.495	.785	.850	1.962						
Var9	.454	.443	.783	.730	.434	.483	.694	.866	1.839					
Var10	.697	.469	.605	.645	.355	.547	.757	.821	.979	1.410				
Var11	.329	.491	.313	.121	.160	.157	.249	.411	.393	.595	1.332			
Var12	.436	.448	.301	.370	.227	.300	.287	.678	.395	.630	.550	1.250		
Var13	.353	.731	.444	.464	.068	.579	.474	.500	.088	.060	.220	.402	1.195	
Var14	.316	.360	.287	.292	.318	.091	.249	.122	- .002	.029	.227	.300	.299	1.516

KEY

Var1	Seemed interested in the topic
Var2	Used examples abstract ideas
Var3	Presented the lecture smoothly
Var4	Integrated the material effectively
Var5	Followed an outline
Var6	Followed a logical sequence of thought in his lecture
Var7	Was well prepared
Var8	Acted relaxed
Var9	Spoke clearly and distinctly
Var10	Spoke with vigor and enthusiasm
Var11	Emphasized important points by raising his voice
Var12	Voice was animated
Var13	Examples were presented which were clearly related to the central topic
Var14	Used purposeful non-verbal behavior

Frame of Reference Group Covariance Matrix

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12	Var13	Var14
Var1	.852													
Var2	.424	1.209												
Var3	.533	.760	1.280											
Var4	.398	.555	.880	1.196										
Var5	.311	.593	.693	.728	1.352									
Var6	.428	.682	.827	.708	.956	1.279								
Var7	.482	.521	.787	.6854	.800	.811	1.551							
Var8	.436	.571	.667	.366	.362	.457	.567	1.220						
Var9	.334	.632	.680	.589	.483	.532	.659	.826	1.375					
Var10	.555	.285	.533	.430	.332	.432	.562	.608	.545	.863				
Var11	.328	.315	.387	.202	.257	.205	.299	.528	.564	.418	.836			
Var12	.428	.432	.573	.322	.283	.332	.445	.346	.308	.385	.297	1.241		
Var13	.527	.849	.747	.574	.468	.579	.576	.453	.527	.406	.429	.567	1.344	
Var14	.485	.615	.773	.622	.627	.475	.583	.535	.584	.511	.466	.771	.845	1.619

KEY

Var1	Seemed interested in the topic
Var2	Used examples abstract ideas
Var3	Presented the lecture smoothly
Var4	Integrated the material effectively
Var5	Followed an outline
Var6	Followed a logical sequence of thought in his lecture
Var7	Was well prepared
Var8	Acted relaxed
Var9	Spoke clearly and distinctly
Var10	Spoke with vigor and enthusiasm
Var11	Emphasized important points by raising his voice
Var12	Voice was animated
Var13	Examples were presented which were clearly related to the central topic
Var14	Used purposeful non-verbal behavior

Comparison Group Covariance Matrix

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12	Var13	Var14
Var1	1.647													
Var2	1.041	1.539												
Var3	1.076	.978	1.855											
Var4	1.060	1.151	1.169	1.522										
Var5	.663	.698	.945	.926	1.548									
Var6	.823	898	.865	.980	1.109	1.472								
Var7	1.169	.967	1.263	1.234	1.027	1.148	2.344							
Var8	.942	.698	1.245	1.071	.750	.853	1.568	2.318						
Var9	.796	.789	1.156	.913	.710	.694	1.054	1.266	1.988					
Var10	1.089	.947	.958	.967	.799	.825	1.237	1.055	1.189	1.618				
Var11	1.049	.975	.851	.993	.821	.838	1.171	.948	1.007	1.294	1.690			
Var12	.839	.922	.908	.842	.527	.665	.969	.895	.845	.973	1.201	1.701		
Var13	.837	1.068	.891	1.097	.716	.993	1.119	.757	.889	.948	1.035	.935	1.610	
Var14	.716	.832	1.427	1.055	.733	.839	1.117	1.000	.958	.873	.842	1.221	1.085	2.226

KEY

Var1	Seemed interested in the topic
Var2	Used examples abstract ideas
Var3	Presented the lecture smoothly
Var4	Integrated the material effectively
Var5	Followed an outline
Var6	Followed a logical sequence of thought in his lecture
Var7	Was well prepared
Var8	Acted relaxed
Var9	Spoke clearly and distinctly
Var10	Spoke with vigor and enthusiasm
Var11	Emphasized important points by raising his voice
Var12	Voice was animated
Var13	Examples were presented which were clearly related to the central topic
Var14	Used purposeful non-verbal behavior

REFERENCES

Armenakis, A. A. & Bedeian, A. G. (1982). On the measurement and control of beta change: Reply to Terborg, Maxwell, and Howard. Academy of Management Review, 7, 296-299.

Armenakis, A. A., & Zmud, R. W. (1979). Interpreting the measurement of change in organizational research. Personnel Psychology, 32, 709-723.

Athey, T. R., & McIntyre, R. M. (1987). Effect of Levels-of-Processing theory and social facilitation theory perspectives. Journal of Applied Psychology, 72, 567-572.

Bedeian, A. G., Armenakis A. A.,& Gibson, R. W. (1980). The measurement and control of beta change. Academy of Management Review, 5, 561-566.

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. Journalof Applied Psychology, 63, 301-308.

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

Bernardin, H. J., & Buckley, M. R., (1981). Strategies in rater training. Academy of Management Review, 6, 205-212.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology,

65, 60-66.

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 62, 64-69.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 20, 238-252.

Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 63, 135-144.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.

Costin, F. (1974). Measuring lecturing behavior of college instructors. Professional Psychology, 1, 106-108.

Cronbach, C. J. (1955). Processes affecting scores on understanding of others and assuming "similarity". Psychological Bulletin, 52, 177-193.

Davison, M.L. (1985). Multidimensional scaling. New York:Wiley.

Fisicaro, S. A. (1988). A reexamination of the relation between halo error and accuracy. Journal of Applied Psychology, 73, 239-244.

Golembiewski, R. T. , Billingsley, K. ,& Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated in OD designs. Journal of Applied Behavioral Science, 12, 133-157.

Guion, R. M. (1965). Personnel testing. New York: McGraw-Hill.

Hahn, D. C. & Dipboye, R. L. (1988). Effects of training and information on the accuracy and reliability of job evaluations. Journal of Applied Psychology, 73, 146-153.

Hayduk, L. A . (1987). Structural Equation Modeling with Lisrel. Maryland: The John Hopkins University Press.

Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. Journal of Applied Psychology, 64, 502-508.

Joreskog, K.G., & Sorbom, D. (1985). LISREL VI: Analysis of linear structural relationships by the method of maximum likelihood. Uppsala, Sweden: University of Uppsala.

Joreskog, K.G., & Sorbom, D. (1988). LISREL VII: Analysis of linear structural relationships by the method of maximum likelihood. Uppsala, Sweden: University of Uppsala.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 36, 29-33.

Landy, F.J. (1985). Psychology of work behavior. Homewood, IL.: Dorsey Press.

Landy, F. J., & Farr, J. L. (1980). Performance ratings. Psychological Bulleting, 87, 72-107.

Landy, F. J., & Trumbo, D. A. (1980). The psychology of work behavior (rev. ed.), Homewood, IL.: Dorsey Press.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.

Millsap, R. E. & Hartog, S. B. (1988). Alpha, Beta, Gamma change in evaluation research: A structural equation approach. Journal of Applied Psychology, 73, 574-584.

Muliak, S.A., James, L.R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C.D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105(3), 430-445.

Murphy, K. R., & Balzer, W. K. (1981, August). Rater errors and rating accuracy. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based ratings: Consequences for rating accuracy. Journal of Applied Psychology, 71, 39-44.

Murphy, K. R., Balzer, W. K., Kellam, K. L., & Armstrong, J. G. (1984). Effects of the purpose of rating on accuracy in observing teacher behavior and evaluating teaching performance. Journal of Educational Psychology, 76, 45-54.

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. Journal of Applied Psychology, 67,

320-325.

Pulakos, E. D. (1985). A comparison of rater training programs: Error training and accuracy training. Paper presented at Southeastern Psychological Association Meetings.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. Organizational Behavior and Human Decison Processes, 38, 76-91.

Roach, D. W. & Gupta, N. (1990). Relationships among components of rating accuracy in a realistic setting. Paper presented at Fifth Annual Meeting of the Society for Industrial and Organizational Psychology.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Personnel Bulletin, 88, 413-428.

Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. Multi-variate Behavioral Research, 17, 343-358.

Smith, D. E. (1986). Training programs for performancce appraisal: A review. Academy of Management Review, 11, 22-40.

Smith, P.C., & Kendall, L.M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Smither, J. W., Barry, S. R., & Reilly, R. R. (1989). An investigation of the validity of expert true score

estimates in appraisal research. Journal of Applied Psychology, 74, 143-151.

Spool, M. D. (1978). Training programs for observers of behavior: A review. Personnel Psychology, 31, 853-887.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. Journal of Applied Psychology, 73, 497-506.

Terborg, J. R., Howard, G. S.,& Maxwell, S. E. (1980). Evaluating planned organizational change: A method for assessing alpha, beta, gamma change. Academy of Management Review, 5, 109-121.

Tennis, C. N. (1989). Responses to the alpha, beta, gamma change typology: Cultural resistance to change. Group and Organizational Studies, 14 (2), 134-149.

Terborg, J. R., Maxwell, S. E., & Howard, G. S. (1982) On the measurement and control of beta change: Problems with the Bedeian, Armenakis, & Gibson technique. Academy of Management Review, 7, 292-295.

Van Driel, O.P. (1978) On Various causes of improper solutions in maximum likelihood factor analysis. Psychometrika, 43, 225-243.

Warmke, D. L., & Billings, R. S. (1979). Comparison of training methods for improving the psychometric quality of experimental and administrative performance ratings. Journal of Applied Psychology, 64, 124-131.

Werts, C.E., Rock, D.A., Linn, R. L., & Joreskog, K.G. (1977). Validating psychometric assumptions within and between populations. Educational and

Psychological Measurement, 37, 863-871.

Yukl, G. A. (1981). Leadership in Organizations. Englewood Cliffs, N.J.: Prentice-Hall.

Zmud, R. W. & Armenakis, A. A. (1978). Understanding the measurement of change. Academy of Management Review, 3, 661-669.