AN EMPIRICAL TEST OF THE ASSUMPTIONS OF PROCESSING INVARIANCE IN LABORATORY STUDIES OF PERFORMANCE APPRAISAL

by

Steven E. Walker

Dissertation submitted to the Faculty of the
Virginia Polytechnic Insititute and State Univeristy
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

APPROVED:

Neil M. A. Hauenstein, Chairperson

Danny K. Axsom

Roseanne J. Foti

Robert J. Harvey

Kent B. Monroe

August, 1989

Blacksburg, Virginia

AN EMPIRICAL TEST OF THE ASSUMPTIONS OF PROCESSING INVARIANCE IN LABORATORY STUDIES OF PERFORMANCE APPRAISAL

by

Steven E. Walker

Committee Chairperson: Neil M. A. Hauenstein

Psychology

(ABSTRACT)

Laboratory studies of the cognitive processes of performance appraisal which employ undergraduates as raters necessarily imply one of two assumptions of processing invariance: the constant category assumption or the constant familiarity assumption. The constant category assumption is implied in studies having undergraduate students rate unfamiliar occupations, and then generalizations are made to supervisors in organizational settings who are much more familiar with the job they are rating. On the other hand, the constant familiarity assumption is implied in studies in which generalizations of performance ratings are made only when raters rate an occupation with which they are familiar (i.e., when students rate teaching). The present study tested these two competing assumptions by varying job familiarity and appraisal purpose in a 2 (rater population) X 2 (target occupation) X 2 (appraisal purpose) X 3 (performance) mixed factorial design. 40 professional carpenters recruited from various contracting firms in the Southwest Virginia, and

40 undergraduate college students from Virginia Polytechnic Institute viewed videotaped performances of three carpentry students performing four different woodworking tasks, and three teaching assistants giving brief lectures. Appraisal purpose was manipulated orthogonally by telling half of the subjects to form a general impression of the ratees' performances (impression-set) and telling the other half to remember as many of the behaviors and actions of the subjects as possible (memory-set). Job familiarity served as a repeated measure, and was manipulated by crossing rater population (student vs. carpenter) with target occupation (teaching vs. carpentry). It was predicted that subjects familiar with the occupational category they assessed would (a) vary their processing strategies according to appraisal purpose (i.e., recall more judgments under an impression-set and recall more behaviors under a memory-set); (b) better recall the order of ratee performance information; (c) better discriminate between ratee performance levels; and (d) provide more accurate ratings than unfamiliar raters. Analysis of subjects' free recalls generally failed to support the hypotheses, partly due to the failure of the appraisal purpose manipulation. For outcome measures, results provided partial support for the hypotheses in that job familiarity led to significant differences in performance discrimination and rating accuracy only when subjects rated the carpentry occupation. No differences were seen when subjects rated teachers. While these findings tend to provide greater support for the constant category assumption than for the constant familiarity assumption, some problems with the use and development of the teaching videotapes may have exacerbated these effects. Implications for future performance appraisal research and the application of performance ratings are offered.

ACKNOWLEDGEMENTS

First, and foremost, I would like to express my utmost appreciation to my advisor, Neil Hauenstein. I think Neil would agree that we had (and will continue to have) a very close professional and personal relationship, and without his guidance, this project, along with many other of my recent endeavors, would just not have been. I call Neil the "advisor for all reasons", and to me, this is exactly what he is: a buddy, a mentor, a sage, and now ... a colleague. I truly feel quite fortunate that I had the opportunity to finish the last two years of graduate school under the supervision of a very competent, bright, good friend. Thanks, Neil.

I would also like to express sincere thanks to another very respected professor, Roseanne Foti. Roseanne has pretty much seen me go through the entire process, and I can easily say that I learned more about my field (and its "process") from her than from anyone else. Roseanne is a stellar type as professors, and just nice people go, and I hope that my future experiences in life will allow me to work with more people like her.

To my other committee members, I thank each of you for the time and input into the conduct of my dissertation. Danny Axsom has also become a good friend, and his ability to give insight into my project at a real "down to earth" level was greatly appreciated. R.J. Harvey has become a friend and an important person to me at this time of life change. I thank him for his time, input, and frank advise in helping me make some rather major decisions. Finally, I would like to thank Kent Monroe for his valuable time and effort into making my dissertation a better study. Dr. Monroe was selected to be on my committee because I knew he had a vast

knowledge of research principles that are amenable to any area of the social sciences. I can now say with great confidence that his input during the course of this project only served to improve the quality of my dissertation.

Although this next part will look cluttered, I can honestly say that each and every person listed below has had a significant impact on my personal and/or academic career, and since I don't think one can separate the two, well here she goes...

From the Blacksburg studio, I thank for his music camaraderie, sling-shot sessions, and for being my closest drinking pal over the last four years. Batting second is A cool, down to earth, dependable surf punk who has recently become a very close buddy. Not only that, the guy can cook, install ceiling fans and go to gradual school. Now there's a well-rounded dude! I also want to talk about Here's a guy I'm always dissin' on, but to tell you the truth, deserves it. Naaaa. is a good friend and has also offered me some great perspectives on what school, work, fun, and life is all about. Hey you three guys, you'll be seeing me again... count on it.

Still from the Blacksburg arena, I must mention and Two guys who quickly became good friends, and I'm sorry I got's to leave 'em before we could have hung out some more. Naaaaa, I'm not sorry. How about my old friends here who have seen me grow to be the same obnoxious guy I was when I started: and Thanks for all the fun.

How about the "older" guys who saw me as a mere babe and turned me into the awesome doctor I am today (yea, right): and My role models and pals that I hope to keep in touch with to the end. These guys definitely had a lot to do with the quality of my work and fun in B'burg. and are two other friends that have shaped my Blacksburgian career. Goof, you've been a friend through the whole thing. I'm still trying to figure out what you are, but a friend and "guide" are definitely two of 'em.

I also gots to thank It's awkward, but I cannot conclude this chapter without mentioning the person who may have shaped it the most. More about this on the 11 O'clock news.

Finally, from the Blacksburg boat I must say hello (and good-bye) to We went through the whole shebang together, the only two souls left from our class, and I must say that made it a whole lot easier. Here was a guy that I implicitly trusted and respected like he was my professor. That's why I called him I know someday he is going to be a pretty big wig in that academic domain, but I also know he'll never change. Simply put, a great guy (but an Oriole fan... go figure!). I'll see you again,

From the New York studio, I feel hello's are in order to the unique group of friends I have up North. To the, shall we say, "screwballs" (no offense guys)-- and --how ya' doin'. To the "more image-oriented professional types" (no offense guys)-- (your bagels did contribute to the overall regression equation), (he's a lawyer, too!), and (my internship buddy), thanks for votes of confidence, etc. All the way from F-L-A, thanks to Les for his continuous baseball talks and calls in the middle of the winter telling me how he is hanging out with the windows open while I'm freezing my @#!%* off.

Lastly, I wish to get serious again and thank to the two most profound people in my life, my parents, and All through this process they have given me nothing by support, in every way. They never once questioned either my progress or the decisions I made throughout my stay here at Tech. I have always been given encouragement and guidance from them, and our relationship, even though 500 miles away, has been as close as any child can have with his parents. I hope that this accomplishment, along with what's down the road for me, can help repay all of their kindness and care they have given me over the course of my life. Thanks Mom and Dad.

TABLE OF CONTENTS

Abstract ii

Acknowledgements iv

List of Tables x

Introduction 1

Literature Review 11

Issue of Generalizability 11

Empirical Research on Generalizability 16

Job Familiarity and Performance Appraisal 20

Social Cognition and Job Familiarity 28

Processing Objectives in Social Cognition Research 38

Conclusions and Integration 46

The Measurement of Memory 50

The Measurement of Accuracy 52

Overview of the Study 55

Hypotheses 57

Method 64

Results 79

Summary and Explanations 104

Discussion 114

References 129

Footnotes 140

Tables 141

Appendix A 175

Appendix B 177

Appendix C 179

Appendix D 181

Appendix E 182

Appendix F 184

Appendix G 195

Appendix H 196

Vita 197

LIST OF TABLES

1. Means and Standard Deviations for Manipulation Checks of Appraisal Purpose and Job Familiarity by Appraisal Purpose and Rater Population 141

2. Means and Standard Deviations for Free Recall Measures by Appraisal Purpose and Rater Population Summed Across Performance Level 142

3. MANOVA Table for Number of Behaviors and Number of Judgments by Rater Population, Recall, and Target Occupation for Subjects Receiving a Memory-Set 143

4. ANOVA Table for the Number of Judgments Recalled for the Target Occupation of Teaching 144

5. MANOVA Tables for the Number of Judgments Recalled for each Rater Population Across Target Occupations 145

6. MANOVA Tables for the Number of Behaviors Recalled for each Rater Population Across Target Occupations 146

7. Means and Standard Deviations for Four ARC' Measures of Seriation by Appraisal Purpose and Rater Population 147

8. MANOVA Tables for the Proportion of Judgments Recalled for each Rater Population Across Target Occupations 148

9. MANOVA Tables for the Proportion of Behaviors Recalled for each Rater Population Across Target Occupations 149

10. Means and Standard Deviations for Behavioral Ratings of Average Teaching Performance by Appraisal Purpose and Rater Population 150

11. Means and Standard Deviations for Behavioral Ratings of Good Teaching Performance by Appraisal Purpose and Rater Population 152

12. Means and Standard Deviations for Behavioral Ratings of Poor Teaching Performance by Appraisal Purpose and Rater Population 154

13. Means and Standard Deviations for Behavioral Ratings of Average Carpentry Performance by Appraisal Purpose and Rater Population 156

14. Means and Standard Deviations for Behavioral Ratings of Good Carpentry Performance by Appraisal Purpose and Rater Population 158

15. Means and Standard Deviations for Behavioral Ratings of Poor Carpentry Performance by Appraisal Purpose and Rater Population 160

16. Means and Standard Deviations for Graphic Ratings for Each Level of Teaching Performance by Appraisal Purpose and Rater Population 162

17. Means and Standard Deviations for Graphic Ratings for Each Level of Carpentry Performance by Appraisal Purpose and Rater Population 163

18. Means and Standard Deviations for Summed Behavioral Ratings for both Teaching and Carpentry Occupations by Level of Performance, Purpose, and Rater Population 164

19. Means and Standard Deviations for Summed Graphic Ratings for both Teaching and Carpentry Occupations by Level of Performance, Purpose, and Rater Population 165

20. Means and Standard Deviations for True Scores by Performance Level for Behavioral and Graphic Ratings for the Occupation of Teaching 166

21. Means and Standard Deviations for True Scores by Performance Level for Behavioral and Graphic Ratings for the Occupation of Carpentry 168

22. Means and Standard Deviations for Variance Components of Rating Accuracy for Behavioral and Graphic Ratings of Teaching and Carpentry by Purpose and Rater Population 169

23. Means and Standard Deviations for Correlational components of Rating Accuracy for Behavioral and Graphic Ratings of Teaching and Carpentry by Purpose and Rater Population 170

24. MANOVA Table for Differential Accuracy on the Behavioral Rating Scales by Appraisal Purpose and Rater Population Across Target Occupations 171

25. ANOVA Table for Elevation of Graphic Ratings for Carpentry Performance 172

26. ANOVA Table for Differential Elevation of Graphic Ratings for Carpentry Performance 173

27. ANOVA Table for Differential Accuracy of Graphic Ratings for Carpentry Performance 174

An Empirical Test of the Assumptions of Processing Invariance in Laboratory Studies of Performance Appraisal

Introduction

The shift in research on performance appraisal from rating format to rating process has resulted largely from dissatisfaction with format improvements and rater training to adequately remove error and bias from judgmental ratings (Landy & Farr, 1980; Feldman, 1981; Cooper, 1981). These authors argue that to properly understand and remedy the multiplicity of factors and problems inherent in performance ratings, one must understand the general cognitive processes involved in performance appraisal (Feldman, 1981). Rating *process* is concerned with information search, storage, organization, and recall of ratee behavior. The fundamental assumption underlying recent models of appraisal is that the rater, as a human information processor, has a limited capacity for attention, storage, and selective recall of behavioral events (DeNisi, Cafferty, & Meglino, 1984; Smither, Reilly, & Buda, 1988; Murphy & Balzer, 1986; Mount & Thompson, 1987). As a result, raters engage in *simplification strategies* to attempt to manage the overwhelming amount of information they are faced with in the appraisal context. Through the use of strategies such as classification or categorization (cf. Feldman, 1981), raters can better identify and assign meaning to behaviors of others, infer unobserved attributes, make predictions about the future, and understand the causes of events (Binning, Zaba, & Whatham, 1986).

Recent studies on the cognitive processes of performance appraisal, for the most part, have been conducted in laboratory settings using undergraduate sophomores as raters evaluating teaching performance (Murphy, Blazer, Lockhart, & Eisenman, 1985; Murphy & Balzer, 1986; Murphy, Gannett, Herr, & Chen, 1986; Athey & McIntyre, 1987; Smither et al., 1988; Krzystofiak, Cardy, & Newman, 1988). Yet, other laboratory studies have investigated the rating process using undergraduates as raters evaluating occupations other than teaching (Pulakos, 1984, 1986; Williams, DeNisi, Meglino, & Cafferty, 1984; Williams, DeNisi, Blencoe, & Cafferty, 1985). For example, Williams et al. (1986) had college sophomores rate advanced woodworking students on various dimensions of carpentry performance. An implicit assumption of both teacher evaluation research and other occupation research is that the information processing of college students rating performance in the laboratory is similar to the manner in which supervisors in "real world" organizations process performance information concerning their subordinates. This reliance on undergraduate raters in artificial appraisal contexts has led some researchers (cf. Ilgen & Favero, 1985; Banks & Murphy, 1985) to seriously question the degree to which laboratory studies on performance appraisal are generalizable to real-world, organizational settings.

Ilgen and Favero (1985) cast doubt on the generalizability of laboratory research on person perception and performance appraisal because the main focus of such research, the interaction between raters and ratees, is often missing in lab studies. These authors add that social cognition research in performance appraisal also ignores problems of: (a) observations over time and future rater-ratee interactions; (b) consequences of ratee behavior; and (c) interdependencies between raters and ratees. Banks and Murphy (1985) also address this issue and claim that the recent popularity of cognitive processes in appraisal research will likely widen the gap between research and practice. This is because the aims of practitioners and researchers are often at odds with each other, and social cognition research clearly caters more to those questions that researchers wish to answer.

While Banks and Murphy and Ilgen and Favero focused their criticisms on contextual issues, an equally important issue is whether the processes observed for the typical subject population (undergraduate sophomores) are the same processes that would be seen for a sample of "real world" raters. Hauenstein and Kovach (1988) claim the generalizability problem is not just due to a lack of contextual variables manipulated in the laboratory (i.e., Banks & Murphy, 1985), but also of demonstrating that the information processing strategies of laboratory subjects are similar to organizational raters.

Gordon, Slade, and Schmitt (1986) addressed this question, and concluded that ultimately, generalizability is an empirical question. Thus, the only way to examine external validity is to conduct experiments that contain student and non-student data collected under the same conditions. Gordon et al. (1986) reviewed a large number of studies that had used both student and non-student samples under identical experimental conditions and reported that, in the majority of studies that used statistical tests of between-group differences, results differed between the two samples. Based on this review, Gordon et al. concluded that problems exist in replicating with non-student subjects behavioral phenomena observed in student samples. Further, these authors recommend that researchers be leery of using student samples as subjects in business-related, decision making research.

Greenberg (1987) argues against Gordon et al. (1986) and claims that their conclusions are somewhat premature. Greenberg posits that, like laboratory studies, samples used in field research are also very narrowly defined, and are thus no more generalizable to other organizations than are student samples. Along these lines, Locke (1986) purports that there is no reason to assume that the psychological processes examined in cognitive performance appraisal research are different among student and non-student samples and processes observed in laboratory and field settings are invariant across subject populations. That is, the way students process performance-related information in the laboratory is similar to the way supervisors process such information in organizational settings.

It is apparent that the generalizability issue involving the choice of subjects and setting in appraisal research remains unresolved. Still, many researchers (cf. Murphy et al., 1986; Cafferty, DeNisi, & Williams, 1986; Smither et al., 1988) continue to examine rating processes in the laboratory under the assumption that underlying cognitive processes of raters do not vary across rater populations and thus, findings obtained in the lab represent valid generalizations to real world settings. However, no research to date has directly examined the role of memory and judgment in a performance appraisal task with two distinct subject populations. Like Gordon et al. (1986), Lynch (1983) claims that the issue of generalizability is an empirical question, and researchers can investigate issues of external validity by intentionally including different population groups in their studies. To this end, the primary purpose of the present investigation is to examine the assumption of processing invariance between two distinct, independent rater populations.

The critical issue behind the generalizability of laboratory performance appraisal research is whether the processes used by students to arrive at performance ratings are similar to the processes used by "real world" supervisors. Laboratory research examining processes of performance appraisal is necessarily conducted under the assumption that the rating process is invariant across subject populations. However, the specific nature of the processing invariance assumption depends on the type of study (i.e., rating teacher performance or another occupation). First, for teacher evaluation studies, it is apparent that student raters are familiar with the category of college professor. Thus, the assumption of processing invariance is that students process teaching performance information in the same manner as raters in "real world" occupations process information about their subordinates. We chose to call this assumption of processing invariance the "constant familiarity assumption." Second, for studies in which student raters judge performance in other occupations, the student raters are clearly less familiar with the occupational category than supervisors performing the job in question. Thus, for these studies, the assumption of processing invariance is that unfamiliar student raters process performance information in occupations (other than teaching performance) in the same manner as raters in these occupations judge their subordinates. We chose to label this processing invariance assumption the "constant category assumption."

A key difference between the two assumptions of processing invariance is the role of rater familiarity. In the constant familiarity assumption, high levels of rater familiarity with the category being judged are sufficient to ensure generalizability, and high levels of rater familiarity are more important than the choice of occupational category. In the constant category assumption, rater familiarity is irrelevant to generalizability, and instead, only the choice of a "real world" occupation is important. Fortunately, recent research on rater familiarity provides some insight on the validity of both assumptions.

The results of recent performance appraisal research suggests rater familiarity is a critical variable. For example, Kozlowski and his colleagues have shown that differing levels of job knowledge differentially affect the degree to which raters' conceptual similarity ratings (i.e., their implicit theories of rating covariance) are related to rating covariance and true score covariance (Kozlowski, Kirsch, & Chao, 1986). More specifically, they reported more halo error for raters lacking both job knowledge and familiarity with the ratee. In addition, Hauenstein and Kovach (1988) reported that raters familiar with the interview process recalled more judgments and fewer behaviors of an applicant than unfamiliar raters. Familiar raters were thus seen to use a categorization-based strategy to process applicant information while unfamiliar raters relied on a more behavioral-level strategy. However, familiar raters were seen to be more efficient processors of applicant information because they showed little decrement in behavioral recognition accuracy despite their tendency to use categorization processes (Hauenstein & Kovach, 1988).

These studies show that a rater's level of job familiarity differentially influences the processing of performance information. However, those studies that adhere to the constant category assumption of processing invariance fail to consider the role of job familiarity in performance appraisal. For example, Williams et al. claim that their results "...have several implications for the practitioner and rater training programs" (p. 194). Unfortunately, these authors fail to consider that, in all likelihood, business students' ratings of carpentry performance may not be the same as those ratings made by more familiar, "real world" supervisors of carpenters. Although Williams et al. (1986) generalized their findings to "practitioners", a proper test of the constant category assumption was never conducted. In order to adequately test this assumption of processing invariance, studies like Williams et al. should include samples of raters familiar with the occupation they are rating. These studies may indicate a lack of support for the constant category assumption. In addition, the suggestion that the constant category assumption is untenable is not surprising when examining the social cognition literature.

From the social cognition perspective, the effects of job familiarity on the rating process are not surprising. Fiske and Taylor (1984) suggested familiarity differentially affects cognitive processing such that familiar raters integrate individual pieces of information, which in turn, speeds access, minimizes confusion, and creates larger perceptual units. Furthermore, Markus, Smith, and Moreland (1985) imply that individuals become familiar at judging attributes of others and of themselves that are highly related to their self-concept. Markus et al. reported that individuals who were familiar with the trait of masculinity chunked masculine information more often than persons unfamiliar with this trait. More importantly, however, Markus et al. found that familiar raters varied their processing strategies depending on the stated objectives of the task. That is, when told to form a general impression of a hypothetical ratee, familiar raters unitized information (i.e., formed larger chunks) more than unfamiliar raters. Alternately, when instructed to divide information into detailed action units, familiar raters sorted ratee information into smaller segments than unfamiliar raters. Unfamiliar raters, on the other hand, did not vary their processing strategies according to changes in processing instructions (Markus et al., 1985). Markus et al. concluded that those persons familiar with a given category were able to form inferences about ratees based on both the contextual cues given to them and their familiarity with the stimulus domain of interest.

As implied by Markus et al. (1985), variations in raters' processing of performance information should be dependent upon the raters' familiarity with the occupation being judged. In turn, persons with varying degrees of job familiarity may process, recall, and evaluate others differently depending on the stated purpose for appraisal. Therefore, varying the processing objectives of raters in a performance appraisal task may provide the means for demonstrating the differential effects of rater familiarity on the rating process.

Summary

Essentially, all laboratory research on performance appraisal either makes the constant familiarity assumption or the constant category assumption. Studies using college students to rate teaching performance make the constant familiarity assumption, whereas studies using college students to rate other occupations make the more tenuous constant category assumption. Research on job familiarity (Gordon et al., 1986; Hauenstein & Kovach, 1988) suggests the constant category assumption may be unwarranted. Differences observed in studies comparing student raters to non-student raters may be due to differences in job familiarity, and thus, previous generalizations made from investigations using students to rate occupations other than teaching may be misleading. Given the research on job familiarity, it appears the constant familiarity assumption may be more tenable. To be sure, an adequate test of both assumptions of processing invariance necessarily requires two separate subject populations, each representing a distinct target occupation (Lynch, 1983).

In the current study, two samples of rater populations were recruited. Each sample represented a specific occupational category, and thus, were familiar with their respective job. One group consisted of students, a distinct group of persons likely to be familiar with the occupation of teaching. The other sample comprised carpenters, a group who are just as likely to be familiar with the occupation of carpentry. Job familiarity was then manipulated by having both groups of subjects evaluate performance for both familiar and unfamiliar jobs. This procedure allowed for testing whether raters familiar with one job processed performance information similar to raters familiar with another job, and whether raters unfamiliar with a job processed information similar to or different than raters familiar with that job. In this context, then, job familiarity is a necessary vehicle for testing the assumptions of rating processing invariance. That is, differential levels of job familiarity may moderate the extent to which cognitive processes observed with a student population generalize to a non-student (working) population.

In addition, appraisal purpose is seen as a vehicle for demonstrating the effects of job familiarity. The results of Markus et al. (1985) suggest that observational purpose should only influence those raters familiar with a job. For example, under instructions to form a general impression of ratee performance, raters unfamiliar with a job may still encode, and subsequently recall more ratee behaviors than judgments. Appraisal purpose may then provide a further test of the constant familiarity assumption. Specifically, if only familiar raters change their processing strategies according to changes in processing objectives, then the constant familiarity assumption will be supported. Therefore, by varying levels of job familiarity within raters, and at the same time, manipulating appraisal purpose between raters, a sufficient test of processing invariance will be conducted. In sum, since the primary aim of this investigation is to examine the issue of generalizability of rating processes, a review of this literature is to follow. Following this section, reviews on job familiarity and observational purpose, in both the applied and social cognition literatures will be presented.

To test the assumptions of processing invariance, both carpenters and students will be asked to view videotapes of both teaching and carpentery performance. In this way, two distinct rater populations will provide performance ratings for occupations for which they are both familiar and unfamiliar. Support for the constant familiarity assumption will be seen if: (a) students rating teachers exhibit similar processes as carpenters rating carpentry; and (b) the processes of students rating carpenters and carpenters rating teachers are different from processes exhibited when rating performance with which they are familiar. Alternatively, support for the constant category assumption will be evidenced if differential levels of job familiarity do not moderate the effects of appraisal purpose on the rating process. Markus et al.'s (1985) findings support this contention. That is, the constant category assumption will be supported if: (a) appraisal purpose influences the processes of students rating carpentry in the same way as the processes of carpenters rating carpentry; and (b) if appraisal purpose influences the processes of carpenters rating teaching in the same way as the processes of students rating teaching.

Literature Review

Issue of Generalizability

As indicated above, a number of researchers have discussed the issue of whether findings obtained in a typical laboratory performance appraisal study generalize to organizational settings (cf. Ilgen & Favero, 1985; Banks & Murphy, 1985). In addition, other researchers have more closely examined the generalizability issue by dealing with the nature of the sample employed in laboratory and field studies (e.g., Gordon et al., 1986; Greenberg, 1987). Essentially, this debate concerns the notion of external validity. That is, how generalizable is the causal relationship between X and Y across persons, settings, and times?

Cook and Campbell (1979) claim that external validity refers to the generalizability of a relationship beyond the circumstances under which it is observed by the scientist. For the purposes of the present investigation, we are interested in the generalization of the rating process *across rater populations*. Cook and Campbell (1979) distinguish between generalizing across populations and generalizing to targeted populations.

Briefly, since characteristics of subjects may vary from sample to sample, generalizing across populations is more difficult than generalizing to target populations. In other words, generalizing across subject populations logically presupposes being able to generalize to target populations. However, generalizing to target populations does not necessarily imply generalization across populations. Thus, one possible threat to external validity may be the nature of the persons sampled. The question then is in which categories of persons can a cause-effect relationship be generalized? According to Cook and Campbell, this threat is considered to be a "treatment by selection" interaction such that the type of subjects sampled systematically alters the influence of any manipulations (treatment). Rosenthal and Rosnow (1975, 1984), for example, have shown that subjects who volunteer to serve as research participants are more sensitive and accommodating to certain coercive task-orienting cues than are nonvolunteer or captive participants.

One of the central concerns of the present research is whether the cognitive processes observed in laboratory studies of performance appraisal are generalizable to organizational settings. However, the two assumptions of processing invariance make different predictions regarding the generalizability of laboratory performance appraisal studies. On the one hand, the constant familiarity assumption suggests that levels of job familiarity moderate the degree to which lab studies on performance appraisal are generalizable to "real world" settings. The processes students use to rate teachers should be equivalent to the processes managers in the field use to rate their subordinates. On the other hand, the constant category assumption claims that laboratory studies involving student ratings of occupations other than teaching are generalizable to organizational settings, regardless of the rater's level of job familiarity. Several authors have recently commented on the degree to which laboratory studies on performance appraisal are generalizable to "real world" settings. A review of their conclusions is to follow.

As indicated above, Ilgen and Favero (1985) and Banks and Murphy (1985) cast doubt on the generalizability of laboratory performance appraisal research. Essentially, these authors argue that most laboratory research conducted on performance appraisal ignores many of the contextual variables inherent in organizational settings (i.e., rater-ratee interactions, distinctions between relevant and irrelevant ratee behavior, consequences of ratee behavior, and raters' willingness to rate). While pertinent to the issue of generalizability, the arguments presented by Ilgen and Favero and Banks and Murphy do not directly address the issue of generalizing processes across subject populations. This latter notion is the primary focus of the present investigation. That is, are the cognitive processes observed with student raters, excluding the issue of setting, generalizable across all kinds of raters?

On the one hand, the constant familiarity assumption argues against the pure generalizability of student ratings to ratings of other occupations. Rather, differential levels of job familiarity determine the degree to which ratings made for one occupation are generalizable to another occupation. For example, students' ratings of teachers should be comparable to carpenters' ratings of carpenters, but students' ratings of carpenters should not be comparable to carpenters' ratings of carpenters. On the other hand, the constant category assumption argues for the pure generalizability of student ratings to ratings of other occupations. For this latter assumption, raters' levels of job familiarity are unimportant, only the choice of a "real world" occupation is necessary for addressing the generalizability issue.

Hauenstein and Kovach (1988) considered this point and claim that the generalizability problem is not just due to a lack of contextual variables manipulated in the laboratory, but also of demonstrating that information processing strategies of laboratory subjects are equivalent to real organizational raters. Several researchers have addressed this issue of generalization across student and non-student raters, and this literature will be reviewed below.

Gordon, Slade, and Schmitt (1986) reviewed 32 studies that had used student and non-student samples and reported that, in the majority of studies that used statistical tests of between-group differences (i.e., quantitative studies), results differed between the two samples. On the other hand, for those studies that did not include significance tests of between-group differences (i.e, qualitative studies), the data supported similarity between student and non-student samples. Based on their quantitative findings, Gordon et al. (1986) concluded that the bulk of between-group differences observed among students and non-students (i.e., managers) could be due to differences in of <u>task familiarity</u>. For example, these authors claim that differences between students and managers in a performance appraisal study could be due to differential familiarity with the rating instrument. In addition, Gordon et al. argue that task familiarity may explain the differences obtained between their analyses of quantitative and qualitative studies. Specifically, they found that in four of the seven instances in which no qualitative differences were observed between students and non-students, both groups were equally familiar with the experimental task. In conclusion, Gordon et al. (1986) claim that differential subject familiarity with the experimental task can be used to account for the majority of differences between students and non-students in quantitative investigations.

In response to Gordon et al. (1986), and other researchers claiming nonequivalence of student and non-student samples, Greenberg (1987) argues that student samples are just as generalizable as non-student samples. His major thesis is that, like laboratory studies, samples used in field research are also very narrowly defined, and are thus no more generalizable to other organizations as are student samples. In addition, Greenberg claims that the purpose of any laboratory study using college subjects is not to explain all organizational phenomena, but rather, such research may be a valuable source of insight into some psychological _processes_ operating therein. Locke (1986) agrees with this notion and adds that there is little or no basis to assume these processes are dependent on the samples employed. Further, Locke claims that the similarities between students and non-students are greater than the differences, and any differences cannot be determined deductively. Bernstein, Hakel, and Harlan (1975) and Landy and Bates (1973) claim that limited generalizability cannot be assumed _a priori_ for a given phenomenon. Researchers need to examine issues of external validity empirically via field experiments with two distinct populations. Finally, Lynch (1983) asks researchers to enhance the external validity of their findings by intentionally including various "blocking variables", such as different population groups, in their studies. In this way, using two distinct, homogeneous groups of subjects may provide additional support for a set of theoretical propositions, or may indicate the presence of certain boundary conditions.

Empirical Research on Generalizability

Efforts to empirically examine the assumptions of processing invariance are indeed rare, however, a few studies have indirectly addressed these issues by comparing the validity of job analysis instruments (e.g., DeNisi, Cornelius, & Blencoe, 1987; Friedman & Harvey, 1986). In addition, two studies directly addressed the assumptions of processing invariance by comparing responses of students and managers to similar antecedent conditions (e.g., Neale & Northcraft, 1986; Barr & Hitt, 1986). These investigations also have in common the theme that raters in organizational settings may have more familiarity with their tasks than student raters in the lab. These studies will be briefly discussed below.

DeNisi et al. (1987) examined the content of a job analysis instrument, the Position Analysis Questionnaire (PAQ; McCormick, Jeanneret, & Mecham, 1972), by comparing responses from experienced supervisors and naive students. Previous research had indicated high correlations between PAQ ratings from experts (incumbents and supervisors) and those from naive subjects (students) relying only on job titles or job titles with brief descriptions (Smith & Hakel, 1979). These findings cast doubt on the validity of the PAQ, because high correlations between expert and naive raters imply that the PAQ may only reflect common knowledge about jobs (DeNisi et al., 1987). However, for the purposes of the present research, this study also examines the differences between individuals high in job knowledge and those low in job knowledge on responses to a structured questionnaire. Contrary to the results obtained by previous research, DeNisi et al. showed that the correlations between expert and naive PAQ ratings were significantly different from 1.00, and further, the correlation between experts' and novices' PAQ ratings increased as the number of "does not apply" (DNA) items increased. DeNisi et al. argued that these findings provided enough evidence to conclude that the ratings of naive raters are not correlated with those of expert raters on the PAQ. Along these lines, Friedman and Harvey (1986) found that job-naive raters who were provided with realistic job descriptive information and practice using that information were unable to produce PAQ ratings that correlated highly with those of experts. They concluded that the presence of accurate, readily available descriptions does not enable the job-naive rater to produce the same quality of PAQ ratings as more traditional methods of information collection (Friedman & Harvey, 1986). More important, however, is the suggestion by DeNisi et al. (1987) that the prototypes (i.e., the hypothesized structures of raters that highlight typical features of a job category; Hastie, 1981) of expert raters may be different than those of naive raters. This issue deals directly with the concerns of whether the cognitive processes of raters high in job familiarity are comparable to those low in familiarity.

Neale and Northcraft (1986) compared the influence of framing and performance constraints on the ability of expert and amateur negotiators to reach integrative agreements on a novel task. It was argued that the generalizability of laboratory studies to the "real-world" has received two major criticisms: (1) expert's experience with the decision making process, or (2) expert's familiarity with the decision content should make him or her immune to the biases naive students exhibit in laboratory experiments (Christensen-Szlanski & Beach, 1984; Ebbesen & Konecni, 1980; Hogarth, 1981). Previous research has shown that experts make decisions by recognizing situations as instances of types with which they are familiar and use past successful solutions. Novel situations, however, lead experts to rely on more general principles (Johnson, Duran, Hassebrock, Moller, Prietula, Feltovick, & Swanson, 1981). Neale and Northcraft predicted that as both experts and amateurs gain experience on a novel task, more integrative solutions will emerge, however, experts will reach more integrative solutions sooner than amateurs. Results confirmed the above hypotheses. In addition, the influence of performance constraints and framing biased the decisions for both experts and amateurs. This led to the conclusion that the general rules applied by experts when dealing with novel tasks do not make them immune to decision bias. Thus, support was found for the generalizability of lab decision making studies (Neale & Northcraft, 1986). Neale and Northcraft call for future research to examine whether experts would be less susceptible to decision bias for information related to one's own content domain. That is, are high job knowledge raters more susceptible than low job knowledge raters to varying influences in performance information? This is one of the questions the present investigation will attempt to answer.

Landy and Bates (1973) discovered contrast effects for students making interviewing decisions, but no such effects were found from a professional sample of interviewers. Bernstein et al. (1975) reviewed six studies that had compared the decision processes of employment interviewers and students, and concluded that while students were more lenient than managers, the decision-making processes were judged to be similar between students and managers.

Barr and Hitt (1986) conducted a study of their own to determine the generalizability of students' responses in an interview setting. Using a policy capturing approach, these authors found significantly more explanatory power in expert decision models than in student decision models for perceived favorability and starting salaries of hypothetical job applicants. Managers used fewer and different types of factors in evaluating applicants than students did. Also, students' favorability ratings were higher than managers, and students' recommended starting salaries were higher than those recommended by managers (Barr & Hitt, 1986). Barr and Hitt argued that, unlike students, managers may consciously try to create a selection decision model incorporating a limited subset of available information. Even without conscious processing, managers, through experience, may have developed _a priori_ decision models to use in the selection process. One can further argue that managers, like the familiar raters seen in the Hauenstein and Kovach (1988) study, were using a _categorization_ strategy to enable them to more efficiently process applicant information.

Summary

The literature reviewed above suggests that students and non-students (i.e., managers) may arrive at different decisions and solutions depending on their level of familiarity with a given task. Gordon et al. (1986) claim that the processes used to arrive at performance ratings may be differentially affected by the degree to which raters are familiar with the job they are surveying. The present research wishes to empirically examine whether the rating process is invariant across rater populations, and thus, attempts to address one critical issue of generalizability. A proper test of the assumptions of processing invariance will necessarily involve recruiting samples from two distinct rater populations, each familiar with their respective occupation. In this way, both the constant familiarity and the constant category assumptions can be empirically examined. Therefore, the study of job familiarity seems particularly relevant for the purposes of the present investigation.

The argument that raters high in job familiarity process information differently than raters low in job familiarity is not new. Research in the social cognition literature has shown, for example, that possession of domain-related knowledge facilitates acquisition of future domain-related information, and is thus consistent with the constant familiarity assumption. Other research in person perception implies that raters with highly accessible category constructs process information at a deeper level than persons with less accessible constructs. This, too, is consistent with the constant familiarity assumption. Initially, a review of job familiarity in the performance appraisal literature will be discussed, and then a review of familiarity in the social cognition literature is to follow.

Job Familiarity and Performance Appraisal

As discussed above, a current trend in performance appraisal research concerns the cognitive processes of raters. Since most organizations employ judgmental measures of ratee performance, it is quite appropriate that we study the fundamental processes of raters making performance appraisals. Landy (1987) and Landy and Farr (1980) have urged industrial psychologists to shift their focus from rating formats to rating process in an attempt to more fully comprehend the underlying causes and consequences of rating accuracy and bias. This shift in focus from format to process has led much of the recent performance appraisal research to be based on findings obtained in the social cognition literature. However, research in this domain of information processing tends to focus more on models of performance appraisal and less on the search for individual difference variables that may moderate the relations between appraisal and observation, storage, retrieval and judgment. Job familiarity in the context of performance appraisal has also been borrowed from the social cognition literature. However, in this regard, job familiarity represents one important individual difference variable that should clearly influence the processes of raters in an appraisal context. Therefore, we will consider research in the industrial psychology literature that deals with differences in job familiarity.

Kozlowski, Kirsch, and Chao (1986) examined the effects of job knowledge, ratee familiarity, conceptual similarity and halo error on performance ratings. Subjects were divided into groups of high and low job knowledge and high and low ratee familiarity based on self-reports of the extent to which they were active observers of baseball and how familiar they were with a list of well-known and not-so-well-known ballplayers, respectively. Subjects were then asked to judge the degree of similarity for all possible pairings of seven performance dimensions unique to baseball. Results showed that high job knowledgeable raters were more sensitive to actual performance covariation, whereas low job knowledge raters were more sensitive to their internal conceptual similarity schemas when rating a familiar ratee. Further, raters high in job knowledge tended to rely on actual performance covariation when rating familiar ratees, whereas conceptual similarity ratings were used when rating unfamiliar ratees. Finally, more halo error was seen when raters were either unfamiliar with the content domain or their ratees (Kozlowski et al., 1986).

Kozlowski and Kirsch (1987) conducted a similar study to Kozlowski et al. (1986), however, this time the authors were interested in whether the same effects would occur when ratings were made from memory (i.e., 3 weeks after stimulus presentation). Based on this and the findings of Kozlowski et al. (1986), Kozlowski and Kirsch predicted that when raters have little experience in the relevant job domain they are more apt to rely on their implicit similarities among rating dimension labels to guide the pattern of ratings. This will lead to increased halo and rating inaccuracy over time.

The methodology for the Kozlowski and Kirsch (1987) study was patterned after Kozlowski et al. (1986), except for the new study subjects' ratings were obtained three weeks after exposure to the stimulus materials. Overall, more halo and less accuracy were found for raters low in either job knowledge or ratee familiarity, thus replicating the findings of Kozlowski et al. (1986). Results also indicated weaker conceptual similarity and rating covariation relations, and stronger between rating and true-score covariation profiles for high job knowledge raters. However, raters high in job knowledge but low in rater familiarity had stronger relations between rating covariation and conceptual similarity schemas. Kozlowski and Kirsch propose that knowledgeable raters have better developed prototype systems which enable them to better encode and recall performance-relevant information for specific ratees, and in turn, this leads to easier category assignment of ratee information to proper performance dimensions. Thus, conceptual similarity schemas help high job knowledge raters rate on performance dimensions even when they lack detailed information about their ratees. In contrast, low job knowledge raters lack well-developed prototypes and thus cannot make category assignments when rating unfamiliar ratees; their conceptual similarity schemas will thus not aid in making dimensional ratings.

Smither and Reilly (1989) also examined the relationship between job knowledge and conceptual similarity schemata. More importantly, one experiment (study 3) directly tested the assumptions of processing invariance by having two separate subject populations, professional market researchers (high job knowledge) and undergraduate students (low job knowledge) provide conceptual similarity schemata for the job of market researcher. As expected, Smither and Reilly found that market researchers provided more reliable conceptual similarity schemata than the undergraduate students. These findings were also seen when job knowledge was measured or manipulated within a single rater population (studies 1 & 2). For all three studies, since the high job knowledge raters had more reliable conceptual similarity schemata than the low job knowledge raters, these results tend to support the constant familiarity assumption over the constant category assumption.

Kozlowski and Ford (1988) also examined the effects of familiarity of ratee performance on the rating process. Kozlowski and Ford varied the effects of delay between exposure to prior ratee performance and subsequent performance rating, ratee performance level, rater search constraint, and ratee familiarity on the amount of information acquired about each ratee. Results indicated a significant familiarity by performance level interaction and a delay by familiarity interaction. There was more search for low performing ratees under conditions of high familiarity, and there was more search with less delay for unfamiliar ratees, respectively. There was also a main effect for familiarity such that increases in ratee familiarity led to decreases in search for performance information (Kozlowski & Ford, 1988). Thus, raters accessed more performance information for ratees with whom they had less familiarity. The interesting result here is the delay by familiarity interaction. This finding indicates that there was more search for performance information for raters low in ratee familiarity when the time between exposure to prior information and search was brief. When prior performance information was memory based, raters were less likely to search for information about unfamiliar ratees. Kozlowski and Ford claim that this finding supports the notion that specific ratee information (i.e., ratee familiarity) serves as the basis for a general evaluation or category assignment that is used as the input for a rating judgment.

Hauenstein and Kovach (1988) conducted two studies examining the effects of category familiarity on raters' processing of applicant information in an interview setting. Based on the works of Bargh and Thein (1985) and Markus et al. (1985) [see below], these authors argued that raters familiar with a stimulus domain should be able to efficiently: select ratee behaviors upon which to categorize the ratee, use the selected behaviors to categorize the ratee, and integrate subsequent behaviors into existing categorical judgments (Hauenstein & Kovach, 1988). On the other hand, unfamiliar raters should sample more behaviors because of their uncertainty as to what ratee behaviors are important for impression formation. Also, unfamiliar raters should possess ill-defined categories which result in less categorization of behaviors and less integration of subsequent behaviors into existing categorical judgments. Thus, Hauenstein and Kovach hypothesized that unfamiliar raters should use a behavioral level processing strategy in which memories of ratee performance will be organized around what the ratee looked like and what the ratee did. In contrast, familiar raters should use a categorical processing strategy in which memories will be organized around trait inferences generated from ratee behaviors. In study 1, rater familiarity with the category of interviewing was operationalized as the raters' self-perceived knowledge of the interviewing process. Subjects were first divided into groups of familiar and unfamiliar raters and were then shown videotapes of a staged interview. Subjects then responded to free recall and recognition questionnaires. Overall, results of this correlational study showed that familiar raters recalled fewer specific behaviors, made more dispositional inferences, and were slightly less accurate at recognizing behavioral incidents than unfamiliar raters (Hauenstein & Kovach, 1988). Thus, familiar raters used a categorization process while unfamiliar raters used a behavioral strategy to organize their memories of applicant performance.

Hauenstein and Kovach conducted a second study to examine whether their self-familiarity measure led to a response bias in subjects and that this bias may have accounted for their observed relationships. In their second study, response bias effects were precluded by recruiting subjects who had experience in an interview setting and then having these subjects complete self-reports on the extent to which they were familiar with interviewing. The main hypothesis of study 2 and the methodology were the same as those for study 1. Results corroborated study 1 in that familiar raters tended to use a categorization process to organize their memories of applicant performance whereas unfamiliar raters processed applicant performance at a more behavioral level. This conclusion was supported by the finding that familiar raters recalled significantly more judgments than unfamiliar raters, whereas unfamiliar raters recalled more behaviors than familiar raters. This latter finding, however, was only marginally significant. It should also be noted that while familiar raters used a categorization strategy for the processing of applicant behavior, their recognition accuracy for applicant <u>behaviors</u> did not significantly decline. Thus, familiar raters were found to be more efficient processors of applicant performance (Hauenstein & Kovach, 1988). In sum, these studies demonstrate how rater familiarity with a category <u>moderates</u> how information is processed in an appraisal context.

In a study using undergraduates as raters and football players as ratees, Hauenstein and Walker (1988) examined the effects of judgment complexity and job knowledge on the relationship between memory and accuracy. Hauenstein and Walker hypothesized that when the judgment task is low in compexity, job knowledge will not moderate the memory-appraisal accuracy relationship, and that memory will be independent of appraisal accuracy. However, when the judgment task is high in complexity, the relationship between memory and judgment will vary as a function of job knowledge. Results confirmed their hypotheses. When the judgment task was low in complexity, the correlations between number of behaviors recalled and rating accuracy for high job knowledge and low job knowledge raters was nonsignificant. On the other hand, when the judgment task was high in complexity, the correlations between recall of behaviors and rating accuracy for both types of raters were highly significant, and in the opposite directions. Although only marginally signficant, the same pattern of results were obtained for the number of judgments recalled and rating accuracy between high and low job knowledgeable raters (Hauenstein & Walker, 1988). These authors concluded that job knowledge and judgment complexity moderate the relationship between memory and accuracy of performance appraisal.

Kingstrom and Mainstone (1985) examined the effects of personal acquaintance between raters and ratees and task acquaintance on ratings supervisors gave to their salesforce. Kingstrom and Mainstone postulated that increases in rater task acquaintance (i.e., self-reports of supervisors' familiarity with specific behavioral dimensions or a sales representative's job) lead to increases in personal acquaintance between raters and ratees. In turn, the positive affect resulting from increased personal acquaintance leads to increased rating accuracy. However, such affect may also lead to increased leniency. These authors found that ratees with high personal and task acquaintance with raters had higher sales productivity, had higher (more favorable) overall evaluations, and were more likely to be promoted. Because supervisory ratings were correlated with sales productivity data, they concluded that the relation between task acquaintance and level of ratings could be partially interpreted as familiar raters providing more accurate judgments.

Summary

In sum, the above studies indicate that job familiarity is an important individual difference variable in the performance appraisal context. Kozlowski and his colleagues have consistently shown that job and ratee familiarity influence the search, recall, and judgment accuracy of performance ratings. Hauenstein and Kovach added that category familiarity moderates how information is processed in these settings. Familiar raters tend to use different strategies than unfamiliar raters when making appraisals. Moreover, Hauenstein and Walker showed how job knowledge moderated the relationship between memory and appraisal accuracy. Kingstrom and Mainstone also noted how task acquaintance predicted the level of ratings supervisors gave to their subordinates. Taken together, these studies show how familiarity influences the processing of performance-related information, but more importantly, the results of each of these studies indicate a lack of support for the constant category assumption.

Social Cognition and Job Familiarity

Although the social cognition literature does not discuss job familiarity per se, several studies have been conducted to examine the effects of domain-related knowledge (Chiesi, Spilich, & Voss, 1979), construct accessibility (Higgins, King, and Mavin, 1982; Bargh & Thein, 1985), and self-schema (Markus, 1977; Markus, Smith, & Moreland, 1985) on person perception. The common theme underlying this research is that persons who are familiar with, for example, a construct or a given knowledge domain process and remember information to a greater extent than unfamiliar persons. Also, much of the recent literature on job familiarity in the context of applied psychology has been borrowed from the social cognition literature. Thus, a review of familiarity in the context of social cognition research is to follow.

Domain-Related Knowledge

Chiesi, Splilich, and Voss (1979) conducted five experiments involving how knowledge of a given topic influences the acquisition of topic-related information. The general research design employed was a contrastive one in which two groups of individuals, experts and nonexperts in a given knowledge domain (baseball), were compared on the basis of their acquisition of domain-related information.

For experiment 1, results showed that, as expected, high knowledgeable persons (HK's) were more sensitive to new changes in the game than low knowledgeable persons (LK's). Also, HK's were more sensitive to both important and unimportant changes than LK's, and HK's were relatively more sensitive to changes that were more important to the game. For experiment 2, Chiesi et al. hypothesized that HK's should "unitize" their input information and better recognize a given change as "old" or "new" sooner than LK's. Results indicated that although there were no recognition performance differences between the two groups, HK's did require less information to make their judgments. For these individuals, old information was correctly recognized with less information than new information. For experiment 3, it was predicted that immediate memory for sequences of events should be greater for HK's than for LK's. Again, Chiesi et al. found that support for their hypothesis; HK's demonstrated better recall than LK's for both ordered and scrambled baseball information. Experiment 4 concerned whether HK's were more likely than LK's to generate actions for a given game state that are oriented towards the goals of the game. Results indicated that just this pattern of differences were obtained for HK's and LK's. For experiment 5, Chiesi et al. reported context sentences presented at input increased recall of target information for HK's but decreased recall of target information for LK's. However, there were no differences in recall between HK's and LK's when target information were presented without context.

Taken together, these experiments indicate that knowledge in a given domain facilitates learning (acquisition) of new domain-related information. The acquisition of domain-related knowledge involves mapping input information onto an existing knowledge structure. Thus, knowledge structure differences between high knowledge individuals and low knowledge individuals (i.e., experts and nonexperts) determines acquisition differences. Therefore, this study highlights the fact that knowledge within a given domain influences the amount and type of cognitive processing employed by raters.

Construct Accessibility

Hastie and Kumar (1979) presented subjects with behaviors that were either congruent, incongruent, or neutral with regards to a general, personality impression. Contrary to their expectations, results showed greater recall for behaviors incongruent with subjects' impressions. However, Bargh and Thein (1985) have noted that when memory demands are high, the recall advantage of inconsistent behaviors is diminished (Hastie, 1981; Srull, 1981; Srull, Lichtenstein, & Rothbart, 1985). Bargh and Thein (1985) also speculated that if subjects are instructed to form an impression of ratee performance (i.e., have an *a priori* judgment goal), impressions will be formed during information acquisition. Such impression formation goals have been shown to enhance recall (Lichtenstein & Srull, 1984; Wyer, Srull, & Gordon, 1984; Lingle & Ostrom, 1979). Furthermore, even under conditions where processing demands are high, impression formation will result, and subjects will have better recall for inconsistent information if the to-be-presented information matches the dimensions of social information in which the subject has "chronically accessible mental constructs" (Bargh & Thein, 1985, p. 1132).

Construct accessibility refers to the differences among individuals with regards to the set of constructs they develop and use in order to understand and predict their environments (Kelly, 1955). Higgins, King, and Mavin (1982) reported that subjects were more likely to delete information that was relevant to their inaccessible categories than information relevant to their accessible constructs (Bargh & Thein, 1985). Bargh (1982) has also demonstrated that chronically accessible categories require little attentional resources in order to process relevant stimuli. Thus, in cases where processing demands are great, impressions can still be formed, and recall for inconsistent information is enhanced when raters possess chronically accessible categories. That is, when processing demands are high, prior knowledge of social information should override the limitations in impression formation and recall for incongruent events. Bargh and Thein (1985) tested this hypothesis.

In their study, Bargh and Thein manipulated two levels of information overload (overload vs. nonoverload) and two levels of chronic category accessibility (chronics vs. nonchronics). Based on their responses to a free-response measure of accessible constructs, subjects were either classified as "chronic" or "nonchronic." Bargh and Thein reported that chronic assessors (i.e., subjects familiar with the category of honesty) had greater recall for inconsistent behaviors even in the information overload condition. Nonchronics, however, did not have better recall for inconsistent behaviors in the information overload condition. In addition, while chronics showed a decrease in recall of consistent behaviors under conditions of information overload, these same subjects did not show a decrease in recall of inconsistent behaviors under such conditions. These findings thus support their original hypothesis that prior knowledge of social information should lead to increased recall for incongruent events seen when processing demands are high. Results also indicated that nonchronics in the nonoverload condition had increased recall of inconsistent behaviors, had "on-line processing" (i.e., formed impressions during information acquisition), and took longer to process inconsistent information. That is, they processed inconsistent information at a deeper level than consistent information. Thus, chronic category accessibility, in the form of prior social information, led raters to form impressions of ratees even under conditions where processing demands were great. In the context of the current study, chronic category accessibility is similar to our notion of category familiarity,. This is because individuals' chronically accessible categories are maintained over time, and with constant use, persons are apt to become highly familiar with these categories.

Self-Schema

Many theorists have used the concept of "schema" to explain how persons (i.e., raters) process social information (cf. Bobrow & Norman, 1975; Hastie, 1981). A *schema* is an "abstract, general structure that establishes relations between specific events or entities" (Hastie, 1981; p.41). Schemas have also been defined as "cognitive structures of organized prior knowledge, abstracted from experience with specific instances" (Fiske & Linville, 1980, p.543; in Landman & Manis, 1983). Essentially then, a schema is a data structure for representing the generic concepts stored in memory (Rumelhart, 1984), and they allow individuals to achieve "cognitive economy" by decreasing the amount of information to be processed and stored (Mount & Thompson, 1987).

Markus (1977) claims that the influence of schemas on the selection and organization of information is most apparent when we process information about ourselves. Specifically, when individuals attempt to organize or explain their own behavior in a particular domain, cognitive structures about the self, or self-schemas are formed. Self-schemas are defined as "cognitive generalizations about the self, derived from past experiences, that organize and guide the processing of self-related information contained in an individual's social experiences" (Markus, 1977, p.64).

Markus, Smith, and Moreland (1985) examined the role of the individual's self-concept in social perception. They define a self-concept as "a set of cognitive structures that provide for individual expertise in particular domains of social behavior" (p. 1495). Here, a person's self-concept is conceptualized as a set of self-schemas that organize past experiences and are used to recognize and interpret relevant stimuli in a social setting. Markus et al. use the notion of self-concept to examine the role of familiarity in cognitive processing.

Markus et al. argue that the qualities one uses to characterize oneself become fairly stable parts of his or her behavioral repertoire, and are subsequently used as reference points against which other person's or one's own behavior are judged. Thus, individuals become "experts" at judging those attributes that they consider to be central to their self-concept. Markus et al. have reviewed the literature on the relation between expertise and cognitive processing and report that experts seem to do four things better than novices: (a) recognize when input information is relevant to their domain of expertise and distinguish such information from irrelevant material; (b) "chunk" relevant information into definable and meaningful units and integrate the information with previously acquired information; (c) retrieve that information with greater accuracy and make greater use than the nonexpert of contextual cues to improve the amount of material recalled; and (d) vary the information-processing strategy, depending on the required task. Markus et al. speculate that persons provided with a self-schema in particular domain (hence forth referred to as schematics) should be able to exhibit the many advantages afforded to experts over novices discussed above. Aschematics, on the other hand, are likely to perceive the actions of others only on the basis of the general knowledge shared by all individuals and according to whatever structure can be inferred from the ratee's behavior. This is because aschematics do not possess a relevant self-schema for a given domain, and thus, cannot use their self-concept in processing the behavior of others (Markus et al., 1985).

Markus et al. (1985) hypothesized that schematics as experts in the domain of masculinity, as opposed to aschematics, would be more apt to recognize the relevance of masculine behaviors in a film and see more coherence and meaning among ratee behaviors. This was assumed to be evidenced by schematics forming larger units for masculine-related behaviors than the units formed by aschematics. A second hypothesis predicted that schematics would view ratees as more masculine and more similar to themselves and will have increased memory for ratee behavior.

Results showed that schematics divided the schema-relevant segments into larger units than aschematics, thus providing support for hypothesis 1. Markus et al. interpreted this finding as indicating that schematics used their self-concepts to organize ratee behavior into larger units. Schematics also attributed more masculine attributes to the ratee than aschematics, and saw the ratee as more similar to themselves than aschematics. Contrary to expectations, there were no differences in recognition of schema-relevant or schema-irrelevant behaviors between schematics and aschematics. These results, however, are tempered by the fact that observed differences between schematics and aschematics could have been due to differences other than those revealed by the self-reports of masculinity. The second experiment was also conducted to examine the effects of varying task instructions (i.e., rating purpose) and expertise on raters' processing of ratee behavior. The authors hypothesized that experts would use their self-schemas to organize the ratee's behavior into smaller meaningful units (as opposed to larger units seen in experiment

1) when specific or detailed information is necessary. Aschematics, because they lack the relevant knowledge structures, will rely primarily on the behaviors exhibited by the ratee for understanding and interpretation, and will thus will be unresponsive to changes in rating purpose.

In experiment 2, a different group of subjects were again classified as either masculine-schematic or masculine-aschematic. In addition, subjects in both groups were randomly assigned to one of three rating purpose conditions: a no instruction condition, a detailed condition where subjects were told to divide the films into the "smallest possible action units", and an impression condition in which subjects were given no instructions regarding the size of their units but were told to "pay attention to the actor" because they would be asked to make judgments about him later on. Results supported the hypothesis that schematics would vary their pattern of processing (i.e., unitizing) with variations in rating purpose. Aschematics did not vary their processing strategies, and divided the segments of the films into the same number of units regardless of the requirements of the task. Specifically, schematics unitized more in the no instruction condition, divided information into smaller segments in the detailed condition, and made larger units in the impression formation condition. Aschematics, on the other hand, did not vary the manner in which they organized and interpreted ratee information according to changes in rating purpose. Markus et al. conclude by claiming that schematics, like experts, used relevant contextual information to go beyond the information contained in the films, and that it is their self-schemas that provides the basis for drawing inferences. Conversely, aschematics lack the relevant knowledge structures needed for such interpretation, and must rely on observable behavior when rating ratees.

To summarize, Markus et al.'s experiments showed that expertise within a given knowledge domain led to: (a) increased recognition of schema-relevant information; (b) chunking of stimuli according to task demands; (c) the forming of impressions of meaningful yet impoverished stimuli; and (d) variations in processing strategies depending on the rating purpose. These findings have important implications for the present investigation since raters familiar within a given job domain should theoretically exhibit the same patterns of information processing seen by the "expert" raters in the Markus et al. experiments.

Summary

From the social cognition literature discussed above, it can be concluded that familiarity, or expertise, within a knowledge domain does influence the processing of information. Chiesi et al. (1979) demonstrated that individuals high in knowledge for a given content domain are better at acquiring new, domain-related information than individuals low in domain-related knowledge. Bargh and Thein (1985) showed how raters can use familiarity to enhance information processing even when memory demands were great. Finally, Markus et al. (1985) showed that raters with high levels of expertise varied their processing strategies depending on the processing objectives of the task.

With regards to the present investigation, these studies suggest that raters familiar with a job may process information differently than raters unfamiliar with a job. Further, job familiarity may interact with stated processing objectives to differentially influence recall and performance judgments. For example, raters familiar with a job should recall more judgments than behaviors when the processing objectives involve forming an impression of ratee performance. However, unfamiliar raters should recall more behaviors than judgments even when the processing objectives involve impression formation.

The predictions noted above, as well as the findings of Markus et al. (1985), explicitly consider the role of processing objectives in recall and judgment. Wyer and Srull (1986) and Srull and Wyer (1986) have formulated comprehensive models showing how processing objectives (i.e., goals; purpose) influence cognitive structures and processing of social information. In addition, applied psychologists such as Peters and DeNisi (1988) have also begun to elaborate on the effects of appraisal purpose on performance ratings. Thus, a discussion of processing objectives in social cognition research is to follow.

Processing Objectives in Social Cognition Research

Peters and DeNisi (1988) claim that appraisal purpose may affect raters' ability to give accurate ratings by affecting how raters process information they use for making performance appraisals. Here, purpose equals a "processing objective" (cognitive set), which influences what information is attended to, how that information is stored, and the ease, completeness, and accuracy of its subsequent recall. From this perspective then, raters' motivation to rate accurately is not what is most important (e.g., Zedeck & Cascio, 1982). Rather, raters may process information differently without being completely consciously aware of the differences inherent in opposing processing objectives. Peters and DeNisi argue that viewing appraisal purpose from this perspective provides additional insight into the importance of appraisal purpose, and might provide a more practical link between appraisal purpose and appraisal accuracy.

Appraisal purpose can be varied to provide raters with different processing objectives for a rating task. This line of research stems from recent work seen in the social cognition literature. Cantor, Mischel, and Schwartz (1982) also refer to such research as the effects of perceived goals or purpose on construct accessibility. In the typical experiment, subjects are given information about one or more persons under instructions either to form an impression of the person(s) described (impression-set) or to remember as much information as possible (memory-set). Later, they are asked to recall the information they have received and/or make various judgments about the stimulus persons (Srull & Wyer, 1986). In this regard, impression formation instructions require raters to make judgments, whereas memory-set instructions require raters to merely focus on remembering specific ratee behaviors. Previous research has consistently shown that, even though impression-set subjects had not expected to be given a memory test at the time they received the information, they nevertheless recalled much more of it than memory-set subjects (Wyer & Srull, 1986). This general finding of greater recall under impression-set than under memory-set conditions has been found with recall tasks (Hamilton, Katz, & Leirer, 1980a, 1980b; Srull, 1981, 1983; Srull, Lichtenstein, & Rothbart, 1985; Wyer, Bodenhausen, & Srull, 1984; Wyer & Gordon, 1982) and recognition tasks (Hartwick, 1979; Srull, 1981; Wyer et al., 1984). Although there is not total agreement on this issue, it has been concluded that the recall advantage for impression-set subjects is generally due to these raters actively imposing a subjective organization on the information presented to them (Hastie, Park, & Weber, 1984; Hamilton et al., 1980a, 1980b). Moreover, traits have been shown to serve as retrieval cues under impression formation instructions (Wyer & Gordon, 1982). Memory-set subjects, because of their focus on accuracy of behavioral detail, do not actively integrate and organize information. Rather, subjects receiving a memory-set tend to rely on the distinctiveness of behaviors for subsequent recall (Hamilton et al., 1980a).

Hamilton, Katz, and Leirer (1980a) conducted three different experiments to examine the effects of processing objectives on recall. Hamilton et al. speculated that impression-set instructions lead subjects to actively impose an organization on the information available about a ratee in order to develop a coherent representation of the ratee. Then, encoded information about the ratee becomes organized and represented in memory in terms of a cognitive structure that represents the raters accumulated knowledge about the ratee. It is this cognitive representation that serves as the basis for recall and later judgment in impression formation tasks (Hamilton et al., 1980). However, memory-set instructions emphasize accuracy in recall of individual ratee behaviors and thus do not lead to the integration and organization of newly presented information with previously acquired information. Thus, poorer recall will result under memory-set instructions because of raters' failure to integrate ratee behavior into an organized cognitive structure. For both experiments, results consistently showed increased recall for subjects in the impression-set condition over subjects in the memory-set condition, and furthermore, this difference could not be attributed to item distinctiveness (Hamilton et al., 1980a). The major conclusion regarding increased recall for impression-set subjects over memory-set subjects was that the former group actively integrated and organized the target information into a coherent structure which apparently leads to greater recall.

Hamilton, Katz, and Leirer (1980b) conducted a similar study to the one mentioned above and reported some additional findings. Subjects were asked to read a set of behavior statements that represented four conceptual categories (interpersonal, intellectual, sports, and religious activities). Specifically, when subjects read a list of person's behaviors under memory-set instructions, the order in which they recalled them resembled the order in which the behaviors were presented. In contrast, when they read the behaviors with an impression-set, their recall was clustered in terms of the categories (religious, interpersonal, etc.) the behaviors exemplified. These findings indicate that memory-set subjects may have used some form of rote rehearsal strategy in which individual items were linked together in a way that resembles the temporal order in which they are receive. Impression-set subjects, as in previous research, tend to impose their own organization of the information by linking each item to other items in the same conceptual category. This happened regardless of where in the temporal string the items occur. Once again, the organization used by impression-set subjects led to greater recall (Hamilton et al., 1980b).

Wyer and Gordon (1982) conducted a number of studies to discover what factors lead impression-set subjects to have greater recall than memory-set subjects. Wyer and Gordon manipulated both distinctiveness of behaviors and evaluative and descriptive consistency of behaviors. Descriptive (in)consistency merely deals with whether behaviors are opposite or the same as traits, whereas evaluative (in)consistency concerns the social (un)desirability of behaviors. There were a number of important results reported from these studies. First, although subjects with an impression-set recalled more behaviors than did memory-set subjects, they did not recall more trait adjectives. Second, under memory-set instructions the distinctiveness or unexpectedness of behavioral information was the primary determinant of later recall because recall of behaviors did not reflect organization of traits attributed to the target. That is, memory-set subjects' recall of behaviors did not depend on recall of the trait adjectives. Third, for subjects in the impression-set condition, the recall of a trait adjective facilitated the recall of behaviors that exemplified the trait. Finally, Wyer and Gordon noted that under impression-set instructions, evaluative inconsistency between behaviors and traits led to highest recall. That is, a general instruction to form an impression focused subjects' attention on evaluative (good-bad) dimensions in stimulus information. In sum, recall of behaviors under a memory-set is most affected by the distinctiveness of information, whereas recall of behaviors under an impression-set is influenced by traits serving as retrieval cues for behaviors. Also, the evaluative inconsistency of traits and behaviors increases the probability of recalling specific behaviors (Hastie, Park, & Weber, 1984).

Srull (1983) wanted to see whether the recall advantage for impression-set subjects over memory-set subjects would be maintained under conditions of high information load. Srull predicted that, under low-information-load conditions, impression-set subjects would have greater recall for behaviors than memory-set subjects. However, in cases where situational constraints may influence the number of person categories or number of behaviors per category recalled (i.e., when information is "blocked by person"), Srull hypothesized that impression-set subjects would recall more behaviors per ratee, but would not recall more ratees than memory-set subjects. When information is presented in a random format, impression-set subjects should recall more behaviors per ratee and more ratees than memory-set subjects. Results confirmed the above hypotheses. Even when multiple ratees were used, impression-set subjects had better recall for behaviors than memory-set subjects. Impression-set subjects also recalled more behaviors than memory-set subjects under blocked conditions, and recalled more ratees and behaviors per ratee under random presentation conditions. Srull concluded that impression formation instructions do not allow subjects to access more ratees than memory set instructions, but once a ratee was accessed, many more individual behaviors could be retrieved.

Srull and Wyer (1986) reviewed the literature on how processing objectives act as moderators of the relationship between recall and judgment. Lichtenstein and Srull (1985) postulate that goals act as a moderator in that they often determine whether judgments have been prestored or need to be computed "on-line." Their model asserts that raters instructed to form an impression of a ratee will do so at acquisition (i.e., on-line) and will store that judgment separately and independently from the specific behavioral information that is learned. When the rater is later asked to render judgment, the evaluation already will have been made and will simply be accessed at the time. Thus, under these conditions, a weak relationship between recall of specific behaviors and the global judgment is expected (Srull & Wyer, 1986). On the other hand, raters given no processing objective or given a memory-set will not form a global evaluation at information acquisition. Then, when the rater is later asked to render a judgment, he or she will be forced to retrieve the previously acquired information, and use it as a basis for his or her judgment of the ratee. That is, a judgment will be made on the spot. Therefore, under these conditions, a strong relationship between judgment and the evaluative implications of the recalled information is expected (Srull & Wyer, 1986). Recent research has supported these hypotheses (Sherman, Zehner, Johnson, & Hirt, 1983; Lichtenstein & Srull, 1985).

Foti and Lord (1987) examined the effects of alternative methods of processing information on rating accuracy. These authors postulated that varying the processing objectives of raters would influence the type of schema formed by raters. Two types of schemas were predicted to be seen in their study: person schema and script schema. Person schema involve raters classifying ratees into preexisting categories based on similarity to a stimulus prototype. Script schemas are cognitive structures that describe the appropriate sequence of events in some situation.

Foti and Lord varied three levels of processing objectives (memory-set vs. impression-set vs. realistic portrayal) and two levels of knowledge of task goals (knowledge vs. no knowledge). The authors predicted that subjects receiving a memory set or knowledge of a groups' goals would form script schemas, whereas those receiving an impression-set or no knowledge of goals would form a leader prototype schema. This was because a memory-set induces subjects to remember the temporal order of behaviors observed in a given context, while impression-sets lead persons to focus on related attributes of a target ratee. Results indicated that there was an increased proportion of script information in memory-set and knowledge of goals conditions, and an increased in proportion of leader information recalled in impression-set and no knowledge conditions. However, contrary to expectations, while there was an increased proportion of script items recalled in their temporal order for the knowledge of goals condition, there was no analogous temporal order recall for the memory-set condition. Finally, memory-set and knowledge of goals subjects recalled more event inferences, and impression-set and no knowledge subjects recalled more appearance intrusions indicating script and leader prototype usage, respectively. Thus, their first hypothesis was only partially supported. Overall, however, both observational purpose and knowledge of goals affected the type of schema used to organize and process information. In turn, this influenced the type of information recalled as well as the organization of recalled information (Foti & Lord, 1987).

Foti and Lord suggest that these results mirror the way schemas are used in laboratory and field settings. Specifically, trait-based information dominates in field settings because classification of ratees and prediction of future ratee behavior is most important organizational contexts. This would lead to the predominance of leader prototype schemas in such settings. In contrast, script-based schemas would tend to predominate in laboratory settings because the emphasis here is on storage and remembering ratee information. Further, according to Markus et al. (1985) student raters may lack the expertise or self-schemas necessary to vary their processing according to the stated purposes of appraisal, and thus, are more apt to simply recall behaviors instead of forming impressions (e.g., making evaluations). This outcome may occur even when student raters are explicitly instructed to form an impression of ratee performance. The present research will also address this issue.

Summary

Overall, these studies in the social cognition literature indicate how various processing objectives influence the organization, integration, recall, and subsequent judgments of raters. Generally speaking, impression-sets lead to increased recall over memory-sets, and this is likely due to subjects' imposing an organization of the information presented to them when given an impression-set. In addition, traits serve as retrieval cues for behaviors recalled under such conditions. Also, evaluative inconsistency is another important determinant for enhanced recall of behavioral information. For memory-sets, recall of behaviors is a function of the distinctiveness of such behaviors. Finally, it has been shown that processing goals influence the relationship between recall and judgment, with stronger relations found for impressions formed after information acquisition.

Conclusions and Integration

The above research indicates that processing objectives can affect observer's ratings in a variety of ways. Moreover, for the purposes of the present investigation, examination of the differential effects of processing objectives appears relevant to the study of job familiarity in the appraisal context. For example, Markus et al. (1985) have shown that familiar raters varied their mode of processing depending on the stated objectives of the task. That is, familiar raters recalled more behaviors than traits under a memory-set and more traits than behaviors under an impression-set. Unfamiliar raters, on the other hand, did not vary their processing strategies and merely recalled equal amounts of trait and behavioral information regardless of the processing objective. In sum, the primary role of appraisal purpose in the present study is to highlight the effects of job familiarity on the rating process. While the study of appraisal purpose has important implications in and of itself, purpose is currently viewed as a vehicle for demonstrating the differential effects of familiarity. The findings obtained by Markus et al. (1985) support this contention.

As a relatively unexplored individual difference variable, job familiarity may provide insight into how raters process performance information. More importantly, the influence of job familiarity on performance ratings may shed light on the extent to which the rating process is invariant across subject populations. That is, job familiarity may be one factor that moderates the relationship between performance information and rating outcomes. Research on this factor may help resolve some of the inconsistencies seen in recent performance appraisal literature. Thus, the study of job familiarity may also be seen as a vehicle for examining an even greater issue in appraisal research, namely, whether the rating process is invariant across distinctive subject populations.

The reviews of both the literatures on job familiarity and processing objectives serve to highlight the critical issue of the current investigation the issue of generalizability. Although, not empirically tested together, the literature on job familiarity tends to support the constant familiarity assumption over the constant category assumption. First, homogenous groups of raters have been found to process performance information differently depending on their individual levels of (job) familiarity. Second, processing and outcome differences have been found between students and managers rating occupations only familiar to managers. Third, the interaction of familiarity and appraisal purpose has been shown to differentially affect the nature of social information processing. To be sure, the purpose of the present study is to empirically test both the constant familiarity and constant category assumptions of processing invariance. By directly testing these two processing invariance assumptions within a single study, confident and definitive conclusions can be drawn about the generalizability of laboratory performance appraisal research.

At this point is worthwhile to recall the conclusions made by Gordon et al. (1986). Specifically, these authors speculated that differences observed between student and non-student samples were likely due to differential levels of task familiarity. In those studies that reported significant between-group differences, non-students apparently had more experience, or familiarity, with the rating task. Also, for the majority of those studies not reporting between-group differences, subjects were equally familiar with the tasks of the experiment. Thus, Gordon et al. argue that familiarity moderates the differences observed between student and non-student samples. In addition, Williams et al. (1986) noted that many of their inconclusive findings may have been due to their raters being relatively inexperienced in conducting performance appraisals. Lacking "expert schemas", these raters may not have been able to focus their attention towards relevant performance information, and hence, their recall for such information is diminished. Along these lines, Smither, Reilly, and Buda (1988) claim that job familiarity may have moderated observed relations between ambiguity of ratee performance and subsequent performance ratings. Specifically, they concluded that raters lacking job familiarity may be more likely to produce assimilation effects than raters with high levels of job familiarity. Finally, Barr and Hitt (1986) raise the question of whether it is appropriate to use students as raters for the study of managerial selection decisions. They concluded that managers are more accurate subjects when the purpose of research is to generalize to managerial behavior in organizations. Although the present research does not involve selection decisions per se, results here should shed some additional light on the appropriateness of using student samples in performance appraisal research. Along the lines of Gordon et al. (1986) and Locke (1986), it seems probable that the rating process is consistent across student and worker samples, but only when raters are high in job familiarity. This is consistent with the constant familiarity assumption but not with the constant category assumption.

The present study wishes to examine the effects of appraisal purpose on the rating process with two different subject populations. The majority of research in this area has been confined to the use of student raters, and it is not yet known whether the factors of appraisal purpose that influence the rating process are similar across different types of raters. In addition, no research to date has been conducted on the way appraisal purpose influences raters with varying levels of *job familiarity*. Thus, the present study will examine the effects of job familiarity and appraisal purpose on the rating process with both student and carpenter populations.

The Measurement of Memory

In order to examine the differential effects of job familiarity and processing objectives on the processes raters use to arrive at performance judgments, a free recall task was employed in the current study. This measure was based on subjects' memories of various stimulus information presented to them.

Free recall tasks have frequently been used in social cognition research. This task requires subjects to recall presented information in any order, and is thus completely unstructured by the experimenter. Srull (1984) claims that one advantage of using free recall is that the type and order of items recalled can be used to infer organizational properties of the stimulus information. This may involve categorizing rater recall data for absolute and relative amounts of judgments and behaviors retrieved (i.e., Hauenstein & Kovach, 1988). Free recall measures also seem appropriate for the present study because they typically involve holding stimulus information constant while varying one or more contextual factors (i.e., processing objectives). Further, free recall data can be analyzed for the occurrence of "intrusions." Here, formulas have been developed that correct for guessing on the part of raters.

Free recall data can also be analyzed by measures of category clustering. According to Srull (1984), the order of information recalled by raters may not equal the order of information presented to them. Thus, the degree to which subjects differentially cluster recalled items in conceptually related categories may provide additional insight into how raters organize ratee performance information. Hauenstein and Kovach (1988) suggest a video camera analogy describes the processing of unfamiliar raters. That is, unfamiliar raters may attempt to veridically encode information as it is presented. Thus, it is more likely that unfamiliar raters will recall ratee behaviors in the order with which they are seen, whereas familiar raters not exhibit such ordering effects. To properly assess this and other predictions regarding subjects' order of free recall, a unidirectional measure of seriation, such as the adjusted-ratio-of-clustering prime index (ARC'), will be used (Pelligrino & Battig, 1974). Essentially, the ARC' measure of subjective organization treats the incoming stimulus information as if it were an output sequence (Murphy & Puff, 1982). ARC' provides a scaled estimate of the degree to which an output protocol is organized according to the order of the incoming stimulus information.

In addition to order of recall, several predictions were made concerning the type and amount of information recalled by subjects with differing levels of job familiarity, and also depending on their specific processing objectives. More specifically, it was expected that raters low in job familiarity should recall more behaviors, regardless of processing objectives; raters high in job familiarity should recall more judgments when given an impression-set and recall more behaviors when given a memory-set (cf. Markus et al., 1985). Recall of behaviors and judgments was tabulated according to percentage of recall as well as the proportions of behaviors and judgments recalled.

The Measurement of Accuracy

In addition to using free recall measures, the present study also employed two types of rating scales. First, a behavioral rating scale was used to measure the extent to which each ratee exhibited a number of various job-related behaviors. For each occupation and ratee, subjects rated on 7-point scales the frequency with which they felt the ratee engaged in a certain number of critical incidents (behaviors) embedded in the ratee's performance. Rating scales were anchored by "not at all" to "all of the time." Second, five, 5-point graphic rating scales were used to measure subjects' ratings of each ratees' performance. For each occupation and ratee, ratings were made on four individual performance dimensions as well as an overall rating. Accuracy measures was then used to examine subjects' responses on both behavioral and graphic rating scales.

Accuracy of subjects' ratings was determined by comparing their responses to normative scores or "true scores." True scores for both carpentry and lecture performance were developed by expert raters during pre-testing (i.e., Murphy, Garcia, Kerkar, Martin, and Balzer, 1982). Cronbach's (1955) seminal work on rating accuracy provided the basis for evaluating subjects' ratings. Specifically, Cronbach argued that "any index combining results from heterogenous items presents serious difficulties of interpretation....where possible it should be replaced or extended by separate analyses of judge's ability to predict different qualities of observers" (p. 178). Cronbach claimed that overall accuracy, or D , is the sum of four components: elevation (E), differential elevation (DE), stereotype accuracy (SA), and differential accuracy (DA).

Briefly, according to Murphy et al. (1982), elevation is the average rating, over all ratees and items, given by a rater. The subject whose overall rating is closest to the overall average true score will be the most accurate rater. Differential elevation is the average rating given to each ratee, across all performance dimensions. High DE scores represent accurate rank ordering of ratees on overall performance. Stereotype accuracy is the average rating given to each performance dimension across all ratees. The rater who correctly assesses a groups' relative strengths and weaknesses on individual dimensions will be most accurate. Finally, differential accuracy is the average rating given for each ratee on each individual dimension. DA refers to the rater's ability to distinguish each ratees' performance on each dimension (Murphy et al., 1982). Sulsky and Balzer (1988) note that the primary advantage of using accuracy scores is that they provide a direct, rather than indirect, measure of accuracy. Moreover, unlike traditional rater error measures, accuracy scores make no assumptions regarding the actual distribution of ratee performance.

It is important to note that, except for E, there are two components for each of the three remaining accuracy components. DE, SA, and DA each have variance (standard deviation) and correlational components making up their respective accuracy measures. For example, the correlation component of DE (i.e., DECOR) represents the rater's ability to judge which ratees rate highest on the elevation scale. Similarly, SACOR is the rater's ability to predict the shape and scatter of individual dimensions across ratees. Finally, DACOR is the rater's ability to judge which ratees have the highest scores on each dimension; there is one correlation for each dimension (Cronbach, 1955).

A number of researchers, however, caution the use of these correlational measures of accuracy (cf. Sulsky & Balzer, 1988; Fisicaro, 1988; Becker & Cardy, 1986). For example, Sulsky and Balzer claim that correlational measures do not qualify as accuracy measures, per se, because they do not measure the distance between observed and true score ratings. Fisicaro (1988) reanalyzed two separate data sets and found a correlational accuracy measure used by Borman (1977, 1979) to yield different results than a traditional variance measure of DA. Finally, Becker and Cardy (1986) argue that the correlational components of accuracy measures (DECOR, SACOR, DACOR) yield information that does not equal variance (aggregate) measures of DE, SA, and DA, respectively. They recommend that the choice of accuracy measures will depend on the whether absolute magnitude of error variance in ratings is of concern, separate from the relative importance of error variance in observed score variance (Becker & Cardy, 1986). The implication is that researchers should report both components of accuracy, unless it is known that observed score variance is equal across all raters.

As indicated above, the use of accuracy measures necessarily requires the calculation of true scores. However, Sulsky and Balzer (1988) note a number of serious methodological and theoretical problems with the use of true scores in assessing performance accuracy. They begin by arguing that since there are so many kinds of accuracy scores available, and since most accuracy measures fail to correlate highly with each other, researchers must choose accuracy scores carefully because a given study may yield different results depending on the particular measure(s) chosen. Second, Sulsky and Balzer claim that caution must be used when employing correlational measures of accuracy (i.e., Borman's DA index; Cronbach's correlational indices of accuracy), because these measures are not sensitive to the distances between subject and true score ratings. Thus, standard deviation or variance formulas are preferred. Third, these authors assert that there is a lack of congruence between the various methods of deriving true scores, and as such, true scores are theoretical constructs that can only be estimated and not obtained. Fourth, Sulsky and Balzer argue that procedures for the development of true scores are fallible in that they do not factor in the extent to which expert raters disagree on true scores during pre-testing. Sulsky and Balzer recommend calibrating expert raters with respect to rating formats and each other via a strategy similar to frame-of-reference training (Bernardin & Buckley, 1981).

In light of these limitations, the present study calculated all four measures of accuracy, E, DE, SA, and DA, and both correlational and variance components were determined. However, the variance components of Cronbach's (1955) formulas are preferred. With regards to the use of true scores, pre-testing with carpenters was done to obtain norms for the carpentry tapes, and trained graduate students were used for the development of norms for the lecture tapes. Details about the development of true scores are be presented below.

Overview of the Study

Videotaped performances of three male carpenters performing four tasks, taken from Williams et al. (1986), were presented to two distinct subject populations. Subjects also viewed videotaped performances of three males lecturing on hunger across four performance dimensions. The lecture tapes were similar to those used by Murphy et al. (1985), and were developed by Kevin Murphy (1988). Subjects were male carpenters recruited from various contracting firms in Southwestern Virginia and undergraduates were recruited from Virginia Tech. All subjects viewed both videotapes such that each group was considered to have high job familiarity in their current educational or job status. Familiarity was determined by the manipulation of two independent variables, rater population and target population. Two job knowledge tests, one for carpentry and one for teaching, were developed in order to empirically demonstrate the manipulation of job familiarity. In addition, subjects were also asked to indicate whether they had any past or current experience with the job domain opposite to their own. For example, any carpenters who had college experience, or any students with professional and/or extracurricular experience in carpentry (i.e., construction) were omitted from the study. Half of the subjects in each population were instructed to view performances under an impression set and the other half were instructed to view performances under a memory set. Subjects then provided free recall data, and rated the performance of the ratees.

Hypotheses

Based on the literature reviewed above on information processing and performance appraisal several predictions were derived concerning the effects of job familiarity and appraisal purpose. It should be noted that all hypotheses treat subjects who represent congruent rater populations and target occupations as raters "familiar" with a job, whereas subjects who represent incongruent rater populations and target occupations were referred to as raters "unfamiliar" with a job. The extant literature on appraisal purpose and job familiarity has shown that each variable influences both the processing of subordinate performance-related information and formal appraisal outcomes. These findings suggest different predictions regarding the constant familiarity assumption and the constant category assumption. More specifically, the interaction of job familiarity and appraisal purpose should provide support for the constant familiarity assumption, but not for the constant category assumption.

The next section contains predictions relating to the influence of appraisal purpose and job familiarity on information processing. In the following section, hypotheses were generated concerning the impact of purpose and familiarity on formal outcome variables.

Process Measures

The majority of studies that have varied processing objectives of a task have found superior recall for impression-sets over memory-sets (e.g., Hamilton et al., 1980a, 1980b; Wyer et al., 1984; Srull et al., 1985). Yet, these studies have not distinguished the recall of behaviors from the recall of judgments. Moreover, aside from Markus et al., there has been a failure to consider the combined role of familiarity and purpose on recall. For the current study, this interaction should lead to differences in subjects' recall of performance information. Further, it was predicted that differential levels of job familiarity will moderate the amount and type of information subjects recall. For example, Cohen and Ebbesen (1979) reported that subjects who were told to learn to perform a task while watching an instructional videotape recognized task-related details more accurately than those told to form an impression of the ratee performing the task. Thus, memory-set instructions may lead to improved recall for behavioral details at the expense of other performance information (i.e., performance judgments). Furthermore, this effect may be accentuated by differences in raters' levels of job familiarity.

Markus et al. (1985) found that subjects familiar with a domain varied their processing strategy depending on the stated purpose for appraisal. Unfamiliar raters, on the other hand, recalled equal amounts of traits and behaviors for both impression formation and memory set instructions. Fiske and Taylor (1984) also propose that relative to memorizing, any person-oriented task, especially in familiar situations, may provoke more psychological engagement, deeper processing, and ultimately improved memory. In addition, Hauenstein and Kovach (1988) predicted and found that unfamiliar raters would use a behavioral level strategy when evaluating ratee performance.

These findings indicate that differential levels of job familiarity lead to differences in the processing of performance information. Further, changes in appraisal purpose influence the manner in which familiar and unfamiliar raters process such information. Manipulating job familiarity and appraisal purpose should thus provide an adequate test for the constant familiarity and constant category assumptions. Support for the constant familiarity assumption should be seen if: (a) the processing of performance information is the same for familiar raters across occupations; and (b) the processing of performance information for unfamiliar raters is different from the processing of such information when rating performance with which they are familiar. For example, when rating teachers, students should be sensitive to processing objectives more than carpenters, however, when rating carpenters, carpenters should be sensitive to processing objectives more than students. On the other hand, support for the constant category assumption should be seen if differential levels of job familiarity do not moderate the effects of appraisal purpose on the rating process (i.e., if purpose influences the processes of students rating carpentry the same way as the processes of carpenters rating carpentry). The first set of predictions bear directly on these assumptions.

Hypothesis 1: Raters familiar with a job should vary their processing strategies depending on the stated purpose for appraisal. In contrast, unfamiliar raters should not vary their processing strategies with changes in appraisal purpose.

Hypothesis 1a: Raters receiving memory-set instructions should recall more behaviors than judgments, regardless of their level of job familiarity.

Hypothesis 1b: For recall of judgments, raters receiving an impression-set, relative to raters receiving a memory-set, should recall significantly more judgments when rating an occupation with which they are familiar than when rating an occupation with which they are unfamiliar.

Hypothesis 1c: For recall of behaviors, raters receiving a memory-set, relative to raters receiving an impression-set, should recall significantly more behaviors when rating an occupation with which they are familiar than when rating an occupation with which they are unfamiliar.

Hypothesis 1d: Under memory-set instructions, raters familiar with a job should recall the order of ratee behaviors better than any other raters.

Hypothesis 1e: Raters unfamiliar with a job should recall ratee behaviors in the order with which they are presented. This effect will be seen regardless of the stated purpose for appraisal.

It will be recalled that Markus et at. (1985) found that familiar raters organized schema-relevant material into larger segments than unfamiliar raters. In addition, Hauenstein and Kovach (1988) showed that raters familiar with a category recalled more pure judgments than unfamiliar raters. Familiarity was also negatively correlated with the number of specific behaviors remembered. Moreover, although only marginally significant, unfamiliar raters recalled more behaviors than familiar raters. Therefore, the next set of hypotheses also test the two assumptions of processing invariance. However, it was predicted that the constant familiarity assumption would be supported over the the constant category assumption regardless of the purpose for appraisal. For example, the constant familiarity assumption would be supported if: (a) students recall a greater percentage of judgments than do carpenters when rating teachers, and conversely, carpenters recall a greater percentage of judgments than do students when rating carpenters; and (b) carpenters recall a greater percentage of behaviors than do students when rating teachers, and students recall a greater percentage of behaviors than do carpenters when rating carpenters. However, the constant category assumption would be supported if there are no differences in the percent recall of judgments and behaviors across familiar and unfamiliar occupations. Hypotheses 2 and 3 test these predictions.

> <u>Hypothesis 2:</u> Raters familiar with a job should recall a greater percentage of judgments than raters unfamiliar with a job. This effect should be seen regardless of the stated purpose for appraisal.

> <u>Hypothesis 3:</u> Raters unfamiliar with a job should recall a greater percentage of behaviors than raters familiar with a job. This effect should be seen regardless of the stated purpose for appraisal.

<u>Outcome Measures</u>

The two assumptions of processing invariance were also tested with regards to rating accuracy. The constant familiarity assumption suggests that raters should be better able to discriminate among performance levels of persons within the occupation with which they are familiar. That is, students should be better than carpenters at discriminating among different levels of teaching performance, whereas carpenters should be better than students at discriminating carpentry performance. Kozlowski and

Kirsch (1987), for example, reported significant correlations between raters high in job knowledge and rating accuracy, but only for ratees high in rater familiarity. On the other hand, the constant category assumption predicts that students and carpenters should be equally able to discriminate among contrasting levels of ratee performance in either occupational category. Hypotheses 4, 4a, and 4b test these assumptions.

> Hypothesis 4: Raters familiar with a job should better discriminate between different levels of ratee performance than raters unfamiliar with a job.
>
> Hypothesis 4a: Raters familiar with a job should rate "good" performers significantly higher than "average" performers, and rate "poor" performers significantly lower than "average" performers. Unfamiliar raters, however, should not discriminate between either good and average performance levels or between poor and average performance levels.

Along these lines, predictions were made concerning the effects of job familiarity on the accuracy of behavioral versus judgmental ratings. In the present context, behavioral rating scale items required raters to indicate the frequency with which ratees engaged in various job-related activities. On the other hand, graphic rating scales required subjects to make judgments about ratee performance on a number of dimensions. Specifically, appraisal purpose should interact with job familiarity to differentially influence accuracy of behavioral and judgmental ratings.

Therefore, the two assumptions of processing invariance were tested with regards to dimensional and overall rating accuracy. Consistent with the performance discrimination hypotheses, the constant familiarity assumption implies that raters should be more accurate at rating persons within the occupation with which they are familiar. That is, students should be more accurate than carpenters when rating teaching performance, and carpenters should be more accurate than students when rating carpentry performance. Conversely, the constant category assumption predicts no differences in rating accuracy for students and carpenters, regardless of their respective levels of job familiarity.

In addition, appraisal purpose should interact with job familiarity to differentially influence rating accuracy. For example, under memory-set instructions, students should be more accurate at making behavioral teacher ratings, and carpenters should be more accurate at making behavioral carpentry ratings. Alternatively, under impression-set instructions, students and carpenters should be more accurate at making judgmental ratings for their respective occupations. Both of these predictions are consistent with the constant familiarity assumption and contrary to the constant category assumption. The final set of hypotheses addressed these issues.

<u>Hypothesis 5:</u> Raters familiar with a job should provide more accurate ratings than raters unfamiliar with a job.

<u>Hypothesis 5a:</u> Under memory-set instructions, raters familiar with a job should be more accurate on behavioral ratings than any other raters.

<u>Hypothesis 5b:</u> Under impression-set instructions, raters familiar with a job should be more accurate on judgmental ratings (rating scales) than any other raters.

Method

Subjects

Two samples were used in this study. The first sample consisted of 40 male carpenters recruited from various contracting firms in Southwestern Virginia. These subjects were paid $40 for their participation. Carpenters were chosen because they represent a group of real-world employees, and because validated stimulus tapes of carpentry performance (Williams et al., 1986) were made available to the chief investigator. Data were collected in groups ranging from 1 to 6, and there were 20 carpenters per cell. Experience of carpenters was determined via self-report measures, and all subjects completed a 22-item multiple-choice job knowledge test on carpentry.

For the second sample, 54 undergraduate males attending a large Southeastern university were initially included in the study. These subjects earned two extra credit points towards their introductory psychology grades for participation in the study. These individuals were chosen for inclusion in the present study because they represent persons who were familiar with the domain of teaching performance, and because it is relatively easy to develop stimulus videotapes of teaching performance. Thus, it was assumed that students were quite familiar with the norms of good and poor teaching performance. Further, all subjects were instructed to write down what they felt defined a "good teacher." This was done in order to determine that students were indeed more familiar with teaching than carpenters. Data again was collected in groups ranging from in size from 1 to 6, and there were 20 students per cell.

Students were assumed to be familiar with teaching (lecturing) performance because of their recent and relevant experience in this content domain. Carpenters, on the other hand, were assumed to have less experience in this area because: (a) most carpenters do not have any college education; and (b) carpenters are rarely involved in formal lecture-type situations. Carpenters were assumed to be familiar with carpentry performance by definition. Students, however, were assumed to be unfamiliar with carpentry performance because this job domain is strictly outside the academic setting. To ensure that these conditions were met, students and carpenters were administered two job knowledge tests, one for carpentry and one for teaching. In addition, all subjects received screening measures to determine whether any of the raters had extra experience with the occupation opposite to their own. Examination of students' carpentry test scores led to the exclusion of 14 subjects from further analyses. All students who were dropped had scored no better than 50% correct on the carpentry test. Thus, there were $\underline{n}$=40 students in the final sample.

Stimulus Materials

Performance information was contained on two 1/2-in. color videotape cassettes and was presented to subjects on a 19-in. color television monitor. The carpentry tapes were a subset of those used by Williams et al. (1986) and consisted of twelve individual performance segments showing three workers performing four tasks (sawing, sanding, hammering, and staining). Each ratee's performance segment lasted approximately two minutes, and the total length of the carpentry tape was about ten minutes. The actors were paid, male volunteers recruited from an advanced woodworking class at a technical school in upstate New York (Williams et al., 1986). Three levels of worker proficiency were established. One worker performed 3 out of 4 tasks correctly (75% proficient), one worker performed 2 out of 4 tasks correctly (50% proficient), and one worker performed 1 out of 4 tasks correctly (25% proficient). The videotapes were pretested by Williams et al. (1986) and also passed the manipulation check in that study. In addition, subjects in the Williams et al. study rated the 75% proficient worker significantly higher than both the 50% proficient worker and the 25% proficient worker, but the latter two did not differ significantly. However, these ratings were made two days after observation of ratee performance. For the present study, information was presented blocked by persons. Subjects viewed, in order, a 50% proficient worker, and then either a 25% proficient worker followed by a 75% proficient worker, or a 75% proficient worker followed a 25% proficient worker perform all four woodworking tasks. As in the Williams et al. study, the order in which each worker performed the good and poor tasks was counterbalanced across workers.

The lecture tapes were a subset of tapes obtained from Kevin Murphy (personal communication). These tapes are similar to the ones used by Murphy, Balzer, Lockhart, and Eisenman (1985), and consisted of three different actors lecturing on the topic of hunger. The length of the teaching tapes was approximately the same as the carpentry tapes. Each of the three male lecturer's performances was varied on four teaching dimensions (clarity, organization, responsiveness to questions, and educational value). As with the carpentry tapes, three levels of lecture proficiency were established, good, average, and poor. The tape of the "good" performance contained a lecture delivered by a confident, dynamic teacher who maintained effective rapport with his audience. The "poor" performance lecture showed a different teacher giving the same lecture, but with a hesitant, nervous speaking style, who was ineffective in establishing rapport with the audience. Finally, the "average" lecturer had a mediocre speaking style, which was neither dynamic nor hesitant. This final tape thus portrayed a level of performance which was neither outstandingly good nor bad. As with the carpentry tapes, subjects viewed the average teaching performance first, followed by either the poor (good) teacher, and then the good (poor) teacher.

Design

The overall experimental design of this study was a 2 X 2 X 3 X 2 (rater population x target occupation x performance x purpose) mixed factorial with target occupation, and performance as the within-subjects factors and rater population and appraisal purpose as the between factors. Although subject to empirical verification, job familiarity was represented by congruence between the rater population and target occupation, whereas job unfamiliarity was represented by incongruence between rater population and target occupation. Therefore, all 80 subjects were in both the job familiar and job unfamiliar conditions, whereas only 40 subjects received and impression-set and the other 40 received a memory-set. To produce the 12 cells, appraisal purpose (impression-set vs. memory-set) and rater population (students vs. carpenters) were crossed with target occupation (teaching vs. carpentry), and performance (good, average, and poor). Each between-factor cell contained 40 subjects (20 carpenters and 20 students).

The design of the present study varied depending on the dependent measure employed. For the behavioral rating scales, the design was a 2 X 2 X 3 X 2 mixed factorial. In addition, for outcome measures involving subjects' differentiating between the three levels of ratee performance, the same design was used. However, for the free recall data and accuracy measures, the data was collapsed across all performance levels. Thus, for these variables the design was reduced to a 2 X 2 X 2 mixed factorial with target occupation as the the within factor and appraisal purpose and rater population as the between factors.

Independent Variables

Rater Population. Rater population was manipulated by recruiting two separate samples of subjects, undergraduate students and professional carpenters. Students comprised one independent sample of raters and carpenters made up the second sample of raters. As mentioned above, each rater population was assumed to be distinct from the other, and both populations were equally represented.

Target Occupation. Target occupation was manipulated by separate videotaped performances of both teaching and carpentry. Each target occupation was represented by three distinct ratee performance segments, one set for teaching and one set for carpentry. Thus, teaching comprised one target occupation and carpentry comprised the other occupation.

Job familiarity per se was not experimentally manipulated in this study. However, job famliliarity was measured by administering to both rater populations, two job knowledge tests, one for each target occupation (see below). Job familiarity was also inferred by the manipulation of rater population (students vs. carpenters) and target occupation (teaching vs. carpentry). Subjects were assumed to be "job familiar" if they had congruent rater populations and target populations. On the other hand, subjects were "job unfamiliar" if they had incongruent rater populations and target occupations. More specifically, students rating teaching were assumed to be job familiar, whereas students rating carpentry were assumed to be job unfamiliar. Alternatively, carpenters rating carpentry were assumed to be job familiar and carpenters rating teaching were job unfamiliar. Finally, scores on the respective job knowledge tests were used to confirm the distinction between job familiarity and job unfamiliarity.

Appraisal Purpose. Appraisal purpose was manipulated prior to subjects' viewing the videotapes. For this manipulation, subjects received both written and oral instructions telling them to either "form an impression of the ratees' performance" (impression-set) or to "remember as much information as possible" (memory-set). This manipulation has been extensively used elsewhere (Foti & Lord, 1987; Wyer, Bodenhausen, & Srull, 1984; Hamilton, Katz, & Leirer, 1980a).

Performance. Performance was manipulated by having instances of good, average, and poor performances of both teachers and carpenters. For both sets of stimulus materials, the good performance condition contained 75% proficiency, the average performance condition contained 50% proficiency, and the poor performance condition contained 25% proficiency.

Procedure

For both samples, subjects were brought to a small laboratory classroom at Virginia Tech. Subjects were seated in the classroom and then received instructions as to the purpose of the experiment. All subjects were told that the purpose of the experiment involved how people observe the behavior of workers in different occupations.

Subjects were run in groups ranging in size from 1 to 6. Large groups were not advantageous, because in such instances, all subjects would not have an equal vantage point for the television monitor. After introductions, subjects were administered a consent form and were told by the experimenter that their participation was strictly voluntary (see appendix E for consent forms). Upon completion, the experimenter collected the consent forms and then distributed written instructions prior to observation of the stimulus tapes. Subjects initially completed both job knowledge tests, one for carpentry and one for teaching. After these tests were completed, subjects were told, both in written form and via verbal instructions, that they were going to view two videotapes, one showing carpenters performing woodworking tasks and one tape showing teachers lecturing in a classroom. The order of presentation was counterbalanced so that subjects might view the carpentry tape first or the teaching tape first regardless of the sample being tested.

Before viewing the tapes, half of the subjects were instructed to form an impression of the ratees' performances and the other half were told to simply remember as much ratee information as they can. This constituted the purpose manipulation. Each ratee was given a fictitious name, and subjects were familiarized with the ratees before viewing their performance. This was done by presenting subjects with a still photograph of the ratees' face. After the first tape was completed, paper was be given to the subjects and they were instructed to write down everything they could remember about each of the ratees they had observed. There was one page for each ratee with the ratee's fictitious name on top, and subjects were allowed to use their still photographs to remember which ratee went with which name. This was done because in most appraisal situations, the supervisor is well aware of which subordinate he or she is rating. Subjects were asked to list and number each comment so as to minimize coders' interpretations of subjects' responses. Subjects had 15 minutes to complete the free recall task. After the free recall task was completed, the behavioral and graphic rating scales were distributed, and subjects recorded their responses directly on the dependent measures. These scales were designed to measure accuracy in recall and rating accuracy. Upon completion of the rating scales, subjects viewed the next videotape. Subjects were reminded of their appraisal purpose and began viewing the tape. After the second tape was finished, subjects followed the same procedure they did for the first tape. Counterbalancing of the order of job domains also served to hold constant the likelihood of systematic practice effects resulting from subjects' familiarity with the dependent measures.

After subjects completed the second series of free recall and rating scale measures they were given the opportunity to ask questions about the study and were then debriefed. At the end of the experiment, carpenters were paid $40 and students received two extra credit points towards their introductory psychology grade.

Dependent Measures

Free Recall Measure. Subjects' responses were comprised by a list of comments, numbered from first to last. Subjects were not given any additional instructions on how to respond to this measure (See appendix A for recall sheets for both target occupations). Two independent coders, blind to the experimental hypotheses, analyzed the responses for frequency and type of response emitted by the subjects. Responses were classified as "behaviors", "judgments", or "behaviors tagged with judgments" (cf. Hauenstein & Kovach, 1988). These three categories were intended to represent a continuum from incidents observable on the tapes to statements representing complete conjecture on the part of the raters. Interrater agreement for the 480 free recalls was .70. All discrepancies were resolved by the chief investigator. Subjects responses were also analyzed for seriation (i.e., order effects). Using the ARC' index (Pelligrino & Battig, 1974), subjects' free recalls were examined to see whether they recalled ratee behaviors in the same order in which they were presented.

Behavioral Rating Scales. For each set of tapes, subjects responded to a 13-item behavioral rating questionnaire requiring them to rate the frequency of behavioral incidents emitted by the ratees. All items were written at a high level of specificity. For example, "How often did lecturer #1 read from his notes?" Subjects responded on 7-point scales ranging from "not at all" to "all of the time" (See appendix B for behavioral rating scales). Coefficient alpha for the three, 13-item lecture behavioral rating scales ranged from .64 to .80. Overall, coefficient alpha for the 39-item scale was .70. For the carpentry behavioral rating scales, coefficient alpha's ranged from .63 to .70., and the overall coefficient alpha for the 39-item carpentry scale was .68. Subjects' responses were compared to a set of norms generated for each tape. Accuracy of responses was measured by all four components of accuracy: elevation, differential elevation, stereotype accuracy, and differential accuracy (Cronbach, 1955; Murphy et al., 1982). Both correlational and variance components were calculated (Becker & Cardy, 1986).

Graphic Rating Scales. The teachers' and carpenters' (ratees) performances were evaluated using two sets of five, 5-point graphic rating scales with anchors of poor and excellent. The five scales consisted of the four dimensions embedded in the videotapes and one dimension measuring an overall evaluation (See appendix C for graphic rating scales). Coefficient alpha's for the three, five-item lecture graphic rating scales ranged from .67 to 83. Overall, coefficient alpha for the 15-item scale was .71. For the carpentry graphic ratings, coefficient alpha's ranged from .75 to .88, and the overall reliability of this scale was .79. As with the behavioral rating scales, the accuracy of graphic rating scales were assessed by Cronbach's four components of accuracy. Again, both correlational and variance components were computed.

Manipulation Check. A manipulation check for appraisal purpose was included in the final questionnaire. Subjects' perceptions of the clarity of the purpose, the extent to which they tried to form an impression of the ratees, and the extent to which they tried to remember what the ratees did were assessed by three, 5-point Likert scale items (see appendix D for manipulation check items).

Tests of Job Knowledge. To ensure that the conditions of job familiarity were met, students and carpenters were administered two job knowledge tests, one for carpentry and one for teaching. The carpentry test was composed of 22 multiple-choice items inquiring about different aspects of trim, rough, and finish carpentry. A subject matter expert in carpentry aided in the development of the test items (see appendix F for carpentry test).

Since teaching is not as content-oriented as carpentry, a free-response format was chosen to assess teaching knowledge. Specifically, subjects were instructed to simply write down what they felt were important aspects of good teaching. Both quantity of statements and quantity of critical teaching behaviors, as determined by Harari and Zedeck (1973) were used as dependent variables. Results of both of these tests will be presented below.

Subjects were also given questionnaires asking each group if they have had any experience in carpentry and teaching domains, respectively. All students claiming experience in technical schools or construction work were omitted from the study. Similarly, all carpenters with any recent college experience (i.e., within the past five years) were also excluded from participation. In essence, each of these tests served as manipulation checks for job familiarity.

In addition to the knowledge tests, all subjects responded to a screening measure intended to objectively assess each rater's experience with the target occupation opposite to their own. Several questions were developed to inquire the extent to which each subject has had some full-time, part-time, or extracurricular background with the opposite occupation.

Power Analysis

A power analysis (Cohen, 1977) indicated that, for the detection of large effects, cell sizes of 27 were sufficient to yield a power of .9.

Thus, the sample size of 80 for this 2 X 2 X 3 X 2 (2 within, 2 between) mixed design was deemed appropriate for the current study.

Development of True Scores

Borman (1977) noted that if experts are given enhanced opportunity to examine videotapes of job performance (e.g., multiple exposures to the tapes), the mean rating computed over a number of expert judges provides a "true score" measure of a specific ratee's performance. That is, this mean rating approximates the expected value of the rating obtained from an expert who is observing behavior under optimal conditions (Murphy & Balzer, 1986). Once these true scores are available, accuracy of subjects' ratings of lecturer i on dimension j can be made by comparing their responses with this true score.

For the development of true scores for the three teaching tapes, ten graduate students in psychology served as expert raters. Graduate students were selected because their experience as students and teachers presumably made them quite familiar with the teaching occupation. In accordance with the guidelines set forth by Sulsky and Balzer (1986), the graduate students were first briefed about their task at hand. They were then shown the three videotapes without exposure to the rating scales. After this presentation, the chief investigator initiated conversation about what the group thought were the most salient and relevant dimensions of teaching performance seen on the tapes. This was done in order to establish a common frame of reference among the expert raters. Following this discussion, raters were familiarized with the two rating scales used in the study. Students were also instructed about the different types of rating errors that may bias their ratings (e.g., leniency, severity, halo, central tendency). All questions regarding the clarity of the scales and rating errors were answered. Students then were shown the videotapes a second time with the knowledge that they would have to make both behavioral and judgmental ratings afterwards.

The development of the true scores for carpentry performance followed the same procedure used for the lecture tapes. However, due to scheduling constraints, the subject matter experts for the carpentry tapes were run individually. With this in mind, efforts were taken to ensure that each of the five experts received the same instructions as to how to observe the ratees and use the two rating scales.

In accordance with Borman (1977), a 13 (dimension) X 3 (ratee), Rater X Performance Dimension design was used to determine convergent and discriminant validities of the behavioral ratings, and a 5 (dimension) X 3 (ratee), Rater X Performance Dimension design was used to determine the validities of the graphic ratings. Convergent validity is indicated when experts agree in their ratings, and discriminant validity is seen when experts distinguish between ratees in evaluating each aspect of ratee performance. According to Kavanagh, MacKinney, & Wolins (1971), when ratings are analyzed in a Rater X Performance Dimension design, the intraclass index for the ratee main effect provides a measure of convergent validity, and the intraclass index for the Ratee X Dimension interaction provides a measure of discriminant validity.

For the teaching tapes, the intraclass indices for the ratee effect in the behavior ratings and the judgmental ratings were .60 and .77, respectively. With regards to discriminant validity, the intraclass ratee x dimension effect in the behavior and judgmental ratings were .29 and

.28, respectively. For the carpentry tapes, the intraclass indices for the ratee effect in the behavior ratings and the judgmental ratings were .08 and .33, respectively. With regards to discriminant validity, the intraclass ratee x dimension effect in the behavior and judgmental ratings were .51 and .68, respectively. Although, an unusally low level of convergent validity is seen with the expert carpenters for the behavioral ratings, high levels of discriminant validity on both scales indicate adequate agreement in experts' discriminating among ratees along dimensions. For the lecture tapes, adequate levels of covergent and discriminant validity were demonstrated amongst the expert raters.

Pilot Study

A pilot study was conducted (*n*=8) in order to assess the purpose manipulation. Due to practical constraints, only students were included in this study. Specifically, it could be argued that since target occupation (i.e., subjects' rating both occupations) was a repeated measure, subjects' perceptions of their given observational purpose could change between presentations of the first occupation and the second. Because of the nature of the free recall task, subjects receiving an impression-set were especially vulnerable to such an effect. One way to address this issue was by counterbalancing the order of occupations presented to subjects. Additionally, the pilot study was undertaken to determine whether subjects' free recalls varied depending on their stated observational purpose, and more importantly, whether subjects' free recalls changed because of repeated measures.

The manipulation check item on how clear subjects perceived their observational purpose showed a small difference between purpose conditions (*M*m=4.00; *M*i=3.25). The manipulation check items inquiring about the extent to which subjects formed an impression or memorized details of the ratees' performances were actually opposite of what was expected. Subjects in the impression-set condition (*M*=4.00) reported they were less likely than subjects in the memory-set condition (*M*=4.25) to form an impression of the ratees' performances. Conversely, subjects in the memory-set condition (*M*=3.50) reported they were less likely to memorize details of the ratees' performances than subjects in the impression-set condition (*M*=4.00). These unusual findings led to a modest revision in the wording of the manipulation check items. Specifically, subjects were eventually asked, "To what extent did you TRY to form an impression/memorize details of the ratees' performances." In addition, subjects in the subsequent study were reminded of their observational purpose on two separate occassions.

Further analyses conducted on the pilot data indicated that, across rater populations and target occupations, subjects who received memory-set instructions recalled a much greater proportion of behaviors (*M*=.63) than subjects who received impression-set instructions (*M*=.44). Conversely, subjects receiving impression-set instructions (*M*=.45) recalled a greater proportion of judgments than subjects receiving memory-set instructions (*M*=.33). Although inconsistent with the results of the manipulation check items, these findings suggested that appraisal purpose was effectively varying the types of recall reported by subjects.

Results

Manipulation Checks

Table 1 indicates the means for the three items assessing the appraisal purpose manipulation. Item 1 asked subjects "To what extent was your observational purpose clear?" Responses ranged from 1, "not clear at all" to 5, "very clear." Results of a one-way analysis of variance (ANOVA) showed no significant differences between the two purpose conditions, $\underline{F}(1,78)=2.19$, $\underline{p}>.10$. Thus, subjects in the impression-set condition ($\underline{M}=3.60$) reported approximately the same levels of perceived clarity of purpose as subjects in the memory-set condition ($\underline{M}=3.88$). This null finding was expected since one appraisal purpose should not be perceived as any more or less clear than the other.

Item 2 asked subjects on a 5-point Likert-type scale, "To what extent did you try to form an impression of each of the ratees' performances?" Although the means were in the expected direction, subjects receiving an impression-set ($\underline{M}=4.33$) did not claim to significantly try to form more impressions than subjects receiving a memory-set ($\underline{M}=4.15$), $\underline{F}(1,78)=1.20$, $\underline{p}>.10$. Finally, item 3 asked subjects, on a similar 5-point scale, "To what extent did you try to memorize details of each of the ratees' performances?" The mean responses for this item were identical between appraisal purpose conditions (Impression-set, $\underline{M}=3.93$; Memory-set, $\underline{M}=3.93$). This finding, along with the results of item 2, raised concerns about the effectiveness of the purpose manipulation.

Table 1 also shows the means for three measures used to assess the effectiveness of the job familiarity manipulation. For the occupation of carpentry, a 22-item, multiple-choice job knowledge test was administered to all 80 subjects. Scores were based simply on the number of items answered correctly. Results of this test showed that carpenters had significantly higher scores (M=18.30) than students (M=7.00), $t(1,78)=24.63$, $p<.001$. These findings thus demonstrated that carpenters were more knowledgeable about carpentry than college students. Knowledge of teaching performance was assessed by having all subjects list what they felt were important aspects of good teaching. Subjects' responses were scored in such a way as to yield two measures of teaching knowledge, one quantitative and one qualitative. The quantitative measure merely consisted of summing the total number of responses listed by each subject. Mean comparisons between the two rater populations showed that students listed significantly more comments about good teaching (M=7.03) than carpenters (M=4.70), $t(1,78)=3.90$, $p<.001$. The qualitative measure consisted of tabulating the number of critical teaching behaviors listed by each subject (cf. Harari & Zedeck, 1973). Thus, only those comments which represented an important dimension of teaching performance were included in the analysis. Here, students also reported significantly more qualitative teaching aspects (M=3.30) than carpenters (M=1.85), $t(1,78)=5.58$, $p<.001$. Taken together, these three measures showed that students were more familiar with teaching and carpenters were more familiar with carpentry.

Hypotheses

Process Measures. Hypothesis 1 stated that, depending on their appraisal purpose, raters familiar with a job would vary their processing strategies more than unfamiliar raters. Additionally, hypothesis 1a posited that, regardless of job familiarity, raters receiving memory-set instructions would recall more behaviors than judgments. Hypothesis 1b stated that for recall of judgments, raters receiving an impression-set, relative to raters receiving a memory-set, would recall significantly more judgments when recalling information about a job with which they are familiar than when recalling information about a job with which they are unfamiliar. Hypothesis 1c stated that for recall of behaviors, raters receiving a memory-set, relative to raters receiving an impression-set, would recall significantly more behaviors when recalling information about a job with which they are familiar than when recalling information about a job with which they are unfamiliar. According to these hypotheses, support for the constant familiarity assumption would be demonstrated if: (a) the processing of performance information is the same for familiar raters across occupations; and (b) the processing of performance information for unfamiliar raters is different from the processing of such information when rating performance with which they are familiar. Conversely, support for the constant category assumption would be seen if differential levels of job familiarity do not moderate the effects of appraisal purpose on the rating process.

Table 2 shows the cell means for the total number of behaviors, judgments, and behaviors tagged with judgments recalled by subjects, collapsed over performance level. In order to test hypothesis 1a, a 2 (rater population) X 2 (target occupation) X 2 (recall) repeated measures multivariate analysis of variance (MANOVA), with occupation and recall as the repeated measures, was conducted on the number of behaviors recalled and the number of judgments recalled for both occupations. Since hypothesis 1a specified effects only for subjects receiving memory-set instructions, the MANOVA was performed only on these subjects; impression-set subjects were excluded from the analysis.

Results of the MANOVA indicated no significant main effect for recall. That is, across occupations, subjects receiving memory-set instructions did not recall significantly more behaviors than judgments. However, several other significant findings were obtained. First, a significant main effect for rater population was seen, $F(1,38)=17.35$, $p<.001$. Across both occupations, students ($M=33.75$) recalled more behaviors and judgments than carpenters ($M=25.50$). Second, a significant rater population by recall interaction was obtained, $F(1,38)=4.57$, $p<.05$ (see table 3 for MANOVA table). Follow-up univariate ANOVA's showed a significant main effect for rater population such that, for the occupation of carpentry, students ($M=9.75$) recalled significantly more behaviors than carpenters ($M=4.80$), $F(1,39)=10.97$, $p<.001$ (eta^2=.22). However, unexpectedly, students ($M=7.70$) also recalled more behaviors than carpenters ($M=4.90$) when the target occupation was teaching, $F(1,39)=7.97$, $p<.01$ (eta^2=.18)

Matched-pairs t-tests were also performed on the number of behaviors and the number of judgments recalled by subjects receiving memory-set instructions. Results showed that, across both occupations, subjects actually recalled more judgments ($M=16.05$) than behaviors ($M=13.58$), $t(1,39)=-1.40$, $p>.10$. When the analysis was broken down for the occupation of carpentry, subjects with a memory-set recalled about the same amount of behaviors ($M=7.28$) as judgments ($M=7.20$), $t(1,39)=.05$, $p>.10$. For teaching, subjects actually recalled significantly more judgments

(M=8.85) than behaviors (M=6.30), t(1,39)=-2.48, p<.05. These results thus provide no support for hypothesis 1a, and are in fact, opposite to what was predicted. More specifically, raters receiving memory-set instructions were expected to recall more behaviors than judgments, regardless of their level of job familiarity. However, since subjects receiving memory-set instructions failed to recall significantly more behaviors than judgments, hypothesis 1a was not supported.

To test hypothesis 1b, a 2 (rater population) X 2 (appraisal purpose) X 2 (target occupation) MANOVA, with occupation as the repeated measure, was performed on the number of judgments recalled for both carpentry and teaching. Initially, the three-way interaction between rater population X purpose X occupation was nonsignificant, indicating that familiar raters receiving an impression-set did not recall significantly more judgments than unfamiliar raters receiving an impression-set. Results, however, showed a main effect for rater population such that students recalled significantly more judgments (M=8.93) than carpenters (M=6.96), F(1,76)=5.74, p<.05. Also, a significant rater population X purpose interaction was obtained, F(1,76)=4.37, p<.05, as well as a significant rater population X occupation interaction, F(1,76)=6.52, p<.05. Follow-up ANOVA's revealed only a main effect for rater population on the number of judgments recalled for teaching, F(1,76)=14.80, p<.001 (eta^2=.16; see table 4). When viewing the performances of teaching, students recalled significantly more judgments (M=10.73) than carpenters (M=7.33).

To provide a further test of hypothesis 1b, two additional 2 (purpose) X 2 (occupation) repeated measures MANOVA's were performed for each rater population on the number of judgments recalled. For the student population, the MANOVA indicated no significant two-way interaction between purpose and occupation, thus, students who received impression-sets did not recall significantly more judgments when rating a familiar occupation (teaching) than when rating an unfamiliar occupation (carpentry). However, a main effect for target occupation indicated that students recalled significantly more judgments for the occupation of teaching ($\underline{M}$=10.73) than for carpentry ($\underline{M}$=7.13), $\underline{F}$(1,39)=24.53, $\underline{p}$<.001 (see table 5 for MANOVA results). This finding indicates that students recalled more judgments when they assessed a job with which they were familiar (i.e., teaching) than when they assessed a job with which they were unfamiliar.

For the carpenter population, the MANOVA also indicated no significant interaction between purpose and occupation, thus, carpenters who received impression-sets did not recall significantly more judgments when rating a familiar occupation (carpentry) than when rating an unfamiliar occupation (teaching). However, a main effect for appraisal purpose was obtained, $\underline{F}$(1,39)=4.81, $\underline{p}$<.05. Surprisingly, carpenters receiving a memory-set recalled significantly more judgments ($\underline{M}$=7.90) than carpenters receiving an impression-set ($\underline{M}$=6.03). No other significant effects were obtained. Given that the three-way interaction of rater population X purpose X occupation in the overall MANOVA, and the two 2-way interactions of purpose X occupation in the subsequent MANOVA's were all nonsignificant, hypothesis 1b was not supported. Familiar raters receiving impression-set instructions failed to recall significantly more judgments than unfamiliar raters who also received impression-set instructions.

To test hypothesis 1c, a 2 (rater population) X 2 (purpose) X 2 (occupation) repeated measures MANOVA, with occupation as the repeated measure, was performed on the number of behaviors recalled for both occupations. Initially, the three-way interaction of rater population X purpose X occupation was nonsignificant, indicating that familiar raters receiving memory-set did not recall significantly more behaviors than familiar raters receiving a memory-set. However, several other significant effects were obtained. First, a significant main effect for rater population emerged for the number of behaviors recalled, $F(1,76)=19.14$, $p<.001$. Across both occupations, students recalled significantly more behaviors ($M=7.53$) than carpenters ($M=4.70$). Second, a main effect for purpose was found, $F(1,76)=4.37$, $p<.05$. Subjects receiving a memory-set recalled significantly more behaviors ($M=6.79$) than subjects receiving an impression-set ($M=5.44$). Finally, a significant main effect for target occupation was also observed on the number of behaviors recalled, $F(1,76)=10.35$, $p<.01$. Across both rater populations, subjects recalled significantly more behaviors when viewing carpentry performance ($M=7.10$) than when viewing teaching performance ($M=5.13$).

To provide a further test of hypothesis 1c, two additional 2 (purpose) X 2 (occupation) repeated measures MANOVA's were performed for each rater population on the number of behaviors recalled. For the student population, the MANOVA indicated no significant two-way interaction between purpose and occupation. Therefore, students receiving memory-sets did not recall significantly more behaviors when rating teaching than when rating carpentry. However, two significant main effects, one for appraisal purpose and one for target occupation were obtained. The main effect for purpose indicated that students receiving a memory-set (M=8.73) recalled more behaviors than students receiving an impression-set (M=6.33), $F(1,38)=5.81$, $p<.05$. The main effect for occupation showed that students recalled more behaviors when viewing carpentry performance (M=9.03) than when viewing teaching performance (M=6.03) $F(1,38)=8.63$, $p<.01$ (see table 6). Thus, students recalled more behaviors when assessing an unfamiliar job (i.e., carpentry) than when assessing a familiar job. For the carpenter population, the MANOVA also indicated no significant interaction between purpose and occupation, thus, carpenters receiving memory-sets did not recall significantly more behaviors when rating carpentry than when rating teaching. As with hypothesis 1b, the absence of a statistically significant interaction of rater population X purpose X occupation in the overall MANOVA, along with nonsignificant 2-way interactions between purpose and occupation in the subsequent MANOVA's, indicate no support for hypothesis 1c. Raters receiving memory-set instructions who were familiar with the job they were rating did not recall significantly more behaviors than their unfamiliar counterparts.

Hypotheses 1d and 1e dealt with the order in which raters recalled ratee performance information. Hypothesis 1d stated that under memory-set conditions, raters familiar with a job would recall the order of ratee behaviors better than any other raters. Hypothesis 1e predicted that, regardless of appraisal purpose, raters unfamiliar with a job would recall ratee behaviors in the order in which they were presented. These predictions were tested using the Adjusted Ratio of Clustering index (ARC'; Pellegrino & Battig, 1974).

Before presenting the results of these analyses, several comments must be made about the use of ARC' in this study. First, because the lecture tapes did not contain discrete, salient points of reference, analyses of teaching using ARC' were omitted from the study. Thus, only the carpentry tapes were analyzed for order of subjects' recall. Second, it was felt that subjects, particularly raters familiar with a job, should not be penalized because their correct recall of ratee performance was evaluative and not strictly behavioral. Thus, judgments dealing with behaviors that were part of the input list were included in the ARC' analysis. Third, Pellegrino and Battig have cautioned users of ARC' about the presence of negative values obtained after computation. This is because negative values are not comparable to positive ARC' scores. Thus, any raters scoring below zero on the ARC' index were given ARC' values of zero (i.e., chance recall). This adjustment was deemed appropriate since these specific recalls did not contain any observed bidirectional pairwise repetitions, the key variable in the ARC' equation.

For the carpentry tapes, it was found that two ratees exhibited nine salient behaviors considered suitable for inclusion in the input list; the third ratee had eight behaviors on the input list. Essentially, ARC' is calculated by coding the subject's free recall into three different categories: the observed number of bidirectional pairwise repetitions, the expected number of bidirectional repetitions, and the maximum number of bidirectional repetitions (see appendix G for the computational formula of ARC').

Table 7 shows the cell means for the four ARC' scores used in this study. ARC' was tabulated for each level of carpentry performance, (good, average, and poor), as well as an additional ARC' computed by taking the average of the three individual ARC' scores. In order to test the hypotheses, two 2 (rater population) X 2 (appraisal purpose) ANOVA's were conducted on the four ARC' scores. Results of these analyses revealed only a significant main effect for appraisal purpose, and only when subjects assessed good carpentry performance, $F(1,76)=10.65$, $p<.01$ ($\text{eta}^2=.12$). Examination of the means indicates that subjects receiving an impression-set actually had higher ARC' scores ($M=.43$) (i.e., recalled ratee behaviors in better order) than subjects receiving a memory-set ($M=.14$). No other significant findings were obtained for ARC', and moreover, the composite ARC' scores were identical between the two rater populations ($Mc=.28$; $Ms=.28$). Thus, no support was found for hypotheses 1d and 1e.

In conclusion, hypothesis 1 predicted that familiar raters would vary their processing strategies according to their appraisal purpose more than unfamiliar raters. However, the null findings obtained for hypotheses 1a, 1b, and 1c fail to support this prediction. Since appraisal purpose and job familiarity had no substantial influence on subjects' recall, it cannot be said that familiar raters processed ratee information very differently than unfamiliar raters.

Hypothesis 2 stated that, across appraisal purpose, raters familiar with a job would recall a greater percentage of judgments than raters unfamiliar with a job. Conversely, hypothesis 3 predicted that, across purpose, raters unfamiliar with a job would recall a greater percentage of behaviors than raters familiar with a job.

Support for hypothesis 2 would be demonstrated if the MANOVA performed on the proportion of judgments recalled led to a significant rater population X target occupation interaction. Although this interaction was nonsignificant, analysis of a 2 (purpose) X 2 (occupation) repeated measures MANOVA conducted separately for each rater population led to a significant main effect for occupation when students were selected as the rater population, $\underline{F}(1,38)=15.03$, $\underline{p}<.001$. Students recalled a greater proportion of judgments when they assessed teaching performance ($\underline{M}=.61$) than when they assessed carpentry performance ($\underline{M}=.42$; see table 8). The same MANOVA conducted on proportion of judgments recalled for carpenters also revealed a significant main effect for occupation, $\underline{F}(1,38)=7.54$, $\underline{p}<.01$. However, examination of the means indicated that, similar to the students, carpenters recalled a greater proportion of judgments for teaching ($\underline{M}=.62$) than for carpentry ($\underline{M}=.48$; see table 8). This finding is contrary to the prediction made by hypothesis 2. Therefore, since both familiar and unfamiliar raters recalled proportionally more judgments for teaching than for carpentry, hypothesis 2 was only partially supported.

Support for hypothesis 3 would be demonstrated if the MANOVA performed on the proportion of subjects' behavioral recall led to a significant interaction between rater population and target occupation. The MANOVA conducted on the proportion of behaviors recalled did provide a significant interaction between rater population and target occupation, $\underline{F}(1,76)=4.68$, $\underline{p}<.05$, and a marginally significant interaction between appraisal purpose and target occupation, $\underline{F}(1,76)=3.83$, $\underline{p}<.10$. Two additional 2 (purpose) X 2 (occupation) repeated measures MANOVA's were performed on each rater population for proportion of behaviors recalled.

For students, the expected main effect for target occupation was significant, $F(1,38)=13.12$, $p<.001$ (see table 9). Students recalled a greater proportion of behaviors for carpentry ($M=.54$) than for teaching ($M=.35$). The MANOVA performed on the carpentry population, however, did not reveal any significant effects for target occupation on the proportion of behaviors recalled (see table 9). Thus, hypothesis 3 received only partial support.

Summary: Process Measures. At the outset, a comment is warranted about the ineffectiveness of the appraisal purpose manipulation. For all intensive purposes, the manipulation of purpose did not significantly affect subjects' recall of ratee performance information. This can be seen in the results of the manipulation check items, as well as in the null findings seen with the process measures hypothesizing main effects or interactions involving appraisal purpose. One possible explanation for these unusual results is that the manipulation of target occupation as a repeated measure may have been weakened the effects of appraisal purpose from trial 1 to trial 2. In order to examine this possibility, the manipulation check items and free recalls were reanalyzed using only those responses from trial 1, thus making target occupation a between factor. Measures from either target occupation obtained during the second viewing of the stimulus materials (trial 2) were thus omitted from these analyses. Unfortunately, results of the analyses using occupation as a between factor were entirely consistent with those performed when using occupation as a repeated measure. Therefore, the manipulation of occupation as a repeated measure could not explain the null findings obtained for appraisal purpose in the present experiment.

Overall, results of analyses performed on the process measures provided only mixed support for the hypotheses. It can be concluded that rater population had a greater impact than appraisal purpose on subjects' free recalls. However, the null findings for hypotheses 1a, 1b, and 1c indicate that familiar raters did not vary their processing strategies any more than unfamiliar raters according to appraisal purpose. Thus, hypothesis 1 was not supported, and further, these results tend to favor the constant category assumption over the constant familiarity assumption. Yet, some support for the constant familiarity was seen when students rated both familiar and unfamiliar occupations. Students' recalls were more behaviorally-based when they rated occupations with which they were unfamiliar, and more judgment-oriented when they rated occupations with which they were familiar.

Results of hypotheses 1d and 1e also favored the constant category assumption over the constant familiarity assumption. There were no consistent effects for job familiarity or appraisal purpose on order of recall. Results of hypotheses 2 and 3 provided support for the constant familiarity assumption when students assessed familiar and unfamiliar occupations, yet supported the constant category assumption when carpenters assessed familiar and unfamiliar occupations. In fact, both groups of raters recalled significantly more judgments when rating teaching than when rating carpentry. However, students did recall a greater proportion of behaviors when rating carpentry than when rating teaching. These findings thus provide some evidence that the processing of performance information was not the same across occupations. Thus, to some extent, job familiarity did influence subject' recalls.

Outcome Measures

The first two hypotheses concerned with subjects' ratings of ratee performance dealt with how well raters discriminated between various performance levels. The second three hypotheses examined how accurate raters were in providing behavioral and judgmental ratings. These hypotheses will be reviewed one at a time; that is, first the results of the performance discrimination hypotheses will be presented followed by results of the rating accuracy hypotheses.

Performance Discrimination. Hypothesis 4 stated that raters familiar with a job would better discriminate between different levels of ratee performance than raters unfamiliar with a job. Hypothesis 4a predicted that raters familiar with a job would rate good performers significantly higher than average performers, and rate poor performers significantly lower than average performers. Unfamiliar raters, however, were not expected to distinguish between either good and average or between poor and average performance levels.

Tables 10, 11, and 12 indicate the cell means for each item on the behavioral rating scales for teaching performance, and tables 13 through 15 show the cell means for the behavioral ratings of carpentry. Tables 16 and 17 show cell means, by item, for the graphic ratings of teaching and carpentry, respectively.

It was felt that presentation of the analyses performed on each of the individual items of the behavioral rating scales would detract from the meaning of the hypotheses. Therefore, analyses were performed on the sums of each of the three, 13-item behavioral rating scales. Prior to doing this, the individual items were examined to ensure that the results of any significant main effects for rater population were in a systematic direction. Analyses performed on each of the 13 items of each of the three behavioral rating scales indicated that the ratings made by the two groups of raters conformed to this criterion. That is, basically all main effects for rater population of each of the 39 items indicated that raters familiar with a job provided systematically higher ratings than raters unfamiliar with a job (higher ratings denote more negative ratee behavior). Therefore, all of the analyses reported below employ the sum of each 13-item behavioral rating scale as the dependent measure. The same strategy was also used for analyses of the graphic rating scales.

Cell means for the summed behavioral and graphic ratings ratings by performance level for both teaching and carpentry occupation are shown in tables 18 and 19, respectively. Initially, two 2 (rater population) X 2 (appraisal purpose) X 3 (performance) MANOVA's with performance as the repeated measure were performed on subjects' summed behavioral ratings for each of the three teachers, and each of the three carpenters. Then, two 2 (rater population) X 2 (purpose) X 3 (performance) MANOVA's with performance as the repeated measure were performed on subjects' summed graphic ratings for each of the three teachers and carpenters, respectively. Results will be presented by occupation, with data from the behavioral and graphic ratings of teaching performance presented first followed by data for the target occupation of carpentry.

For teaching, none of the multivariate main effects for rater population, purpose, or any of their interactions reached appropriate levels of statistical significance. However, a multivariate main effect was found for performance, $\underline{F}(2,75)=259.75$, $\underline{p}<.001$. Matched-pairs t-tests showed that both groups of raters were able to correctly discriminate between good and average, and between poor and average teachers. As expected, students rated the average teacher (M=52.00) significantly higher than the good teacher (M=36.63), t(1,39)=5.83, p<.001; and rated the average teacher significantly lower than the poor teacher (M=70.90), t(1,39)=-11.55 p<.001 (higher scores indicate more negative ratee behavior). Surprisingly however, carpenters rated the average teacher (M=51.05) significantly higher than the good teacher (M=36.38), t(1,39)=6.85, p<.001; and rated the average teacher significantly lower than the poor teacher (M=67.63), t(1,39)=-9.57, p<.001. Although students were expected to discriminate between average and good, and between average and poor teacher performance, carpenters were not. The findings that carpenters were able to effectively discriminate between the various levels of teaching behavior does not support hypotheses 4 and 4a.

The MANOVA performed on the summed, 5-item graphic rating scales for teaching provided results similar to those reported for the behavioral ratings. First, however, a significant multivariate main effect for rater population was obtained, F(1,76)=9.71, p<.01. Second, while no significant rater population X performance interaction was seen, a significant main effect for performance was obtained, F(2,75)=206.83, p<.001. Matched-pairs t-tests performed on the three summed graphic ratings revealed that both groups of ratees correctly discriminated between average and good, and between average and poor performance levels. As expected, students rated the average teacher (M=14.20) significantly lower than the good teacher (M=19.50), t(1,39)=-6.73 p<.001; and rated the average teacher significantly higher than the poor teacher (M=9.50), t(1,39)=7.34

p<.001 (higher scores indicate better teaching performance). Surprisingly however, carpenters rated the average teacher (*M*=15.60) significantly lower than the good teacher (*M*=20.60), *t*(1,39)=-6.33 *p*<.001, and rated the average teacher significantly higher than the poor teacher (*M*=10.93), *t*(1,39)=8.49, *p*<.001. While these findings are not surprising for the student raters, the fact that carpenters aptly discriminated between performance levels again fails to support hypotheses 4 and 4a.

Results of the MANOVA on the summed behavioral ratings for carpentry indicated a significant rater population x performance interaction, *F*(2,75)=12.39, *p*<.001. In addition, a significant multivariate main effect for performance was also obtained, *F*(2,75)=23.85, *p*<.001.

Matched-pairs t-tests performed on the sums of each of the three behavioral ratings actually showed that the carpenters, not the students, failed to discriminate between average and good carpenters. When carpenters were selected as the raters, their mean ratings for the average carpenter (*M*=37.88) were not significantly higher than the good carpenter (*M*=35.83), *t*(1,39)=1.05, *p*>.10. The difference in the means, however, was in the correct direction. The carpenters did correctly discriminate between the average (*M*=37.88) and poor carpenters (*M*=43.28), *t*(1,39)=-2.18, *p*<.05 (higher scores denote more negative behavior). When students were selected as raters, they did discriminate between the average (*M*=27.15) and good carpenters' behaviors (*M*=39.18), *t*(1,39)=-5.63, *p*<.001, but their scores were in the direction opposite to those of the expert raters (rating accuracy scores will be discussed in the next section). Finally, students correctly discriminated between the average (*M*=27.15) and poor (*M*=43.95) carpenters, *t*(1,39)=-7.87, *p*<.001. Thus, although both carpenters and students were able to discriminate between average and poor carpentry behaviors, the fact that students reversed the behavioral ratings of the average and good carpenters provides some support for hypotheses 4 and 4a. Inconsistent with hypothesis 4a, carpenters failed to discriminate between good and average carpentry behavior. However, this finding is not surprising given that the expert raters also failed to distinguish between good and average carpenters on the behavioral rating scales. Analysis of the five expert ratings made on the carpentry behavioral rating scales indicated that these raters rated the average carpenter ($\underline{M}$=29.20) almost the same as the good carpenter ($\underline{M}$=29.00).

The MANOVA performed on the graphic rating scales for the occupation of carpentry resulted in a significant multivariate main effect for rater population, $\underline{F}(1,76)=33.85$, $\underline{p}<.001$, and a significant main effect for performance, $\underline{F}(2,75)=17.42$, $\underline{p}<.001$. Also, a significant multivariate effect was obtained for the rater population x performance interaction, $\underline{F}(2,75)=17.43$, $\underline{p}<.001$.

Matched-pairs t-tests performed on the graphic ratings revealed that both groups of ratees correctly discriminated between average and poor performance levels, however, carpenters failed to distinguish between average and good carpentry performance, and the students rated the average carpenter significantly higher than the good carpenter. More specifically, carpenters rated the average carpenter ($\underline{M}$=14.45) nonsignificantly lower than the good carpenter ($\underline{M}$=15.17), $\underline{t}(1,39)=-1.08$, $\underline{p}>.10$, yet the means were in the expected direction. Carpenters did rate the average carpenter significantly higher than the poor carpenter ($\underline{M}$=12.78),

$t(1,39)=2.25$, $p<.05$ (higher scores indicate better carpentry performance). Surprisingly however, students rated the average carpenter ($M=20.65$) significantly higher than the poor carpenter ($M=15.32$), $t(1,39)=6.33$, $p<.001$; but also erroneously rated the average carpenter significantly higher than the good carpenter ($M=15.05$), $t(1,39)=6.56$, $p<.001$. These results parallel those obtained for the carpentry behavioral ratings, and thus, provide support for hypotheses 4 and 4a. For the occupation of carpentry, then, familiar raters better discriminated between ratee performance levels, and familiar raters better discriminated between good and average, and between poor average performance levels than unfamiliar raters.

Summary: Performance Discrimination. Hypotheses 4 and 4a received no support when the target occupation was teaching. Both familiar and unfamiliar raters adequately discriminated between performance levels on both behavioral and graphic rating scales. Therefore, for teaching, the constant category assumption received greater support than the constant familiarity assumption. It should be noted, however, the means of the carpenters' graphic ratings for teachers were consistently higher than those provided by the student raters. For example, a 2 (rater population) X 2 (purpose) ANOVA performed on the summed graphic ratings indicated a significant main effect for rater population, $F(1,76)=5.52$, $p<.05$ ($eta^2=.07$). Students' summed ratings on the five judgmental teaching dimensions for average teaching performance ($M=9.51$) were significantly lower than the ratings of the carpenters ($M=10.92$). Follow-up univariate ANOVA's conducted on the individual items revealed five significant main effects for rater population, and in all five cases students rated teachers significantly lower than when carpenters rated teachers.

Hypotheses 4 and 4a received some support when the target occupation was carpentry. Although unfamiliar raters distinguished between average and good, and between average and poor performance levels, their ratings between the average and good carpenters were in the wrong direction. Familiar raters, on the other hand, correctly discriminated between average and poor performance levels, but failed to discriminate between average and good carpentry performance. It should be noted, however, that when rating carpenters, the failure of familiar raters to discriminate between average and good carpenters on the behavioral ratings was consistent with the ratings provided by subject matter experts. Furthermore, unfamiliar raters rated the average carpenter significantly lower ($\underline{M}$=27.15) than familiar raters ($\underline{M}$=37.88) when making behavioral ratings, $\underline{F}(1,76)=28.99$ $\underline{p}<.001$ ($eta^2=.26$), and rated the average carpenter significantly higher ($\underline{M}$=20.65) than familiar raters ($\underline{M}$=14.45) when making judgmental ratings, $\underline{F}(1,76)=59.28$, $\underline{p}<.001$ ($eta^2=.24$). These findings thus provide partial support for the constant familiarity assumption such that familiar raters better discriminated between ratee performance levels than unfamiliar raters.

Rating Accuracy. As discussed earlier, analyses of rating accuracy necessarily requires the use of expert ratings or "true scores." Means for the true scores, by dimension and ratee, for the behavioral and graphic ratings of teaching performance are shown in table 20; means for the behavioral graphic ratings of carpentry performance are reported in table 21.

Two forms of rating accuracy measures were computed for the present study, variance and correlational components (see appendix H for computational formulas). Sulsky and Balzer (1986) recommended the use of both of these measures since they demonstrated that the two indices do not necessarily provide the same results. These authors, however, tend to favor the variance measures because these measures actually take into account the distance between subjects' ratings and true score ratings. Thus, results will primarily be presented for the variance estimates of rating accuracy, while any findings gathered from the correlational measures will be presented in brief.

Hypothesis 5 stated that raters familiar with a job would provide more accurate ratings than raters unfamiliar with a job. Thus, a significant main effect for rater population for each occupation is predicted. Specifically, the constant familiarity assumption predicts greater rating accuracy when raters' jobs are congruent with the job they are rating. Additionally, hypotheses 5a and 5b predict interactions of appraisal purpose and job familiarity. Hypothesis 5a stated that under memory-set conditions, raters familiar with a job would have higher rating accuracy on the behavioral rating scales than any other raters, and hypothesis 5b stated that under impression-set conditions, raters familiar with a job would have the highest levels of rating accuracy on graphic rating scales. (i.e., when making judgmental ratings).

Cell means for the variance estimates of the four measures of rating accuracy for both behavioral and graphic rating scales, for both occupations are presented in table 22 (scores closer to zero represent greater accuracy.). Cell means for the correlational estimates of rating accuracy are provided in table 23 (scores closer to 1 represent greater accuracy). Two 2 (rater population) X 2 (appraisal purpose) X 2 (occupation) X 4 (accuracy) MANOVA's, with target occupation and accuracy as the repeated measures, were conducted on the accuracy scores of subjects' behavioral and graphic ratings, respectively. Results of the first analysis on subjects' behavioral ratings showed no significant main effects or interactions of any kind. By itself, this nonsignificant MANOVA fails to support hypothesis 5a. The second analysis performed on subjects' graphic ratings also provided no main effects or interactions involving appraisal purpose. Therefore, hypothesis 5b was not supported.

Since effects for rating accuracy were specified _a priori_, it was deemed appropriate to conduct further analyses on subjects' accuracy scores. Specifically, four 2 (rater population) X 2 (purpose) X 2 (occupation) repeated measures MANOVA's were performed on each of the four variance components of rating accuracy. For example, each MANOVA contained two accuracy scores as the dependent measures, one for each target occupation. Results of these analyses showed no effects for any of the independent variables on elevation, differential elevation, or stereotype accuracy of behavioral ratings. However, a significant main effect for rater population on subjects' differential accuracy scores was obtained, $\underline{F}(1,76)=4.42$, $\underline{p}<.05$ (see table 24). Follow-up univariate ANOVA's on each of the two target occupations indicated only a marginally significant main effect for rater population on differential accuracy scores of behavioral teacher ratings, $\underline{F}(1,76)=3.83$, $\underline{p}<.10$ ($eta^2=.05$). Although not reaching acceptable levels of significance, students were more accurate at discriminating among teachers within dimensions ($M=.94$) than carpenters ($M=1.07$).

The second 2 (rater population) X 2 (purpose) X 2 (occupation) X 4 (accuracy) repeated measures MANOVA conducted on the variance scores of rating accuracy for the graphic ratings of both occupations led to a significant main effect for rater population, $F(1,76)=16.28$, $p<.001$, and a significant rater population X accuracy interaction, $F(3,74)=7.05$, $p<.001$. The overall MANOVA also indicated a significant rater population X occupation interaction, $F(1,76)=16.02$, $p<.01$, and a significant effect for the three-way interaction of rater population X accuracy X occupation, $F(3,74)=4.81$, $p<.005$.

Further 2 (rater population) X 2 (purpose) X 2 (occupation) MANOVA's performed on each variance component of rating accuracy revealed a significant main effect for rater population and a significant rater population X occupation interaction for elevation, differential elevation, and differential accuracy. No significant effects were found for stereotype accuracy. For both sets of graphic ratings, there was a main effect for rater population on subjects' elevation scores, $F(1,76)=10.23$, $p<.005$, and a significant rater population X occupation interaction, $F(1,76)=4.54$, $p<.05$. Also for the graphic ratings, there was a significant main effect for rater population on subjects' differential elevation scores, $F(1,76)=10.39$, $p<.005$, and a significant rater population X occupation interaction, $F(1,76)=8.66$, $p<.005$. Finally, there was a significant main effect for rater population on subjects' differential accuracy scores, $F(1,76)=5.38$, $p<.05$, and a significant rater population X occupation interaction, $F(1,76)=7.32$, $p<.01$.

Follow-up univariate ANOVA's revealed three significant main effects for rater population on the graphic ratings for carpentry. For the graphic ratings of teaching, no significant main effects for rater population on rating accuracy were demonstrated. For carpentry, significant main effects were found for elevation, $F(1,76)=12.57$, $p<.001$ ($eta^2=.14$), differential elevation, $F(1,76)=15.43$, $p<.001$ ($eta^2=.167$), and differential accuracy, $F(1,76)=8.56$, $p<.005$ ($eta^2=.10$; see tables 25, 26, & 27). Only stereotype accuracy did not lead to significant differences between rater populations for the graphic ratings of carpentry.

Examination of the mean accuracy scores indicates that, as raters, carpenters provided more accurate ratings across dimensions and ratees ($M=.40$) than students ($M=.67$; elevation). Carpenters were also more accurate at discriminating between carpenters across performance dimensions ($M=.49$) than students ($M=.79$; differential elevation). Finally, carpenters' were more accurate at discriminating among ratees within performance dimensions ($M=1.24$) than students ($M=1.42$; differential accuracy). However, carpenters ($M=.34$) and students ($M=.33$) were about equal in discriminating among dimensions of carpentry performance (stereotype accuracy). These results provide support for hypothesis 5 in that, for judgmental ratings of carpentry, familiar raters were more accurate than unfamiliar raters.

Since there is no correlational accuracy analog for elevation, only differential elevation (DECOR) and differential accuracy (DACOR) could be examined for comparisons of rating accuracy measures on the graphic ratings of carpentry. Of the two significant main effects found using the variance components, only DECOR provided a significant main effect for rater population on accuracy of carpentry performance, _F_(1,76)=5.99, _p_<.05 (eta^2=.073). As with the variance estimate of differential elevation, carpenters (_M_=.39) had higher correlations for discriminating among carpentry performance than students (_M_=.06). There were no significant findings for DACOR on the carpentry performance ratings.

Taken together with the null findings observed for the behavioral ratings for both occupations, and the null findings for the graphic ratings for teaching, the results presented thus far provide only partial support for hypothesis 5, and no support for hypotheses 5a and 5b. Hypothesis 5 was supported only when familiar and unfamiliar raters provided judgmental ratings for the carpentry occupation. Hypotheses 5a and 5b received no support since appraisal purpose failed to interact with job familiarity on any of the components of rating accuracy. These latter findings can be traced back to the failure of appraisal purpose manipulation.

Summary: Rating Accuracy. Overall, the results of the analyses on rating accuracy provided little support for hypothesis 5, and no support for hypotheses 5a and 5b. The only significant effect for job familiarity was seen when raters made evaluative ratings for carpentry performance. Here, raters familiar with a job were more accurate on three of four rating indices than raters unfamiliar with a job. Carpenters were obviously better able to discriminate between various levels of carpentry performance than were students. However, since there were no significant effects for job familiarity on the behavioral ratings for either occupation (albeit the marginal effect for differential accuracy on teaching), it must be concluded that the constant category assumption received more support than the constant familiarity assumption. Furthermore, the absence of any observed interactions between appraisal purpose and job familiarity also provides greater support for the constant category assumption than for the constant familiarity assumption.

Summary and Explanations

Results of the hypotheses for the process measures generally indicated a lack of support for the constant familiarity assumption. Raters familiar with a job did not process performance information differently than raters unfamiliar with a job. More specifically, hypotheses 1a, 1b, and 1c all predicted main effects or interactions involving appraisal purpose on subjects' recall of performance information. Results did not support any of these hypotheses, and the absence of any significant effects for for appraisal purpose were likely due to the failure of the purpose manipulation.[1]

Hypothesis 2 predicted main effects for job familiarity such that raters familiar with a job would recall a greater percentage of judgments than raters unfamiliar with a job. Results showed that students recalled a greater proportion of judgments when they assessed teaching performance than when they assessed carpentry performance, but carpenters also recalled a greater proportion of judgments when they assessed teaching performance than when they assessed carpentry performance. Also, when teaching was examined as the target occupation, it was found that students recalled more judgments than behaviors, and that students recalled more judgments than carpenters. These results suggest that the recall of teaching was primarily judgment-oriented, and since unfamiliar raters recalled more judgments than behaviors, hypothesis 2 was not supported.

Hypothesis 3 predicted that raters unfamiliar with a job would recall a greater percentage of behaviors than raters familiar with a job. However, only mixed support was found for the hypothesis. Specifically, while students recalled a greater proportion of behaviors for carpentry than for teaching, carpenters recalled about an equal number of behaviors for carpentry and teaching. Thus, the findings obtained for the student sample support the prediction that unfamiliar raters' recall will be behaviorally-based. However, the findings obtained for the carpentry sample were not consistent with this prediction.

Results also showed that for the target occupation of carpentry, all subjects recalled significantly more behaviors than judgments. Further, students recalled more behaviors than judgments when assessing carpenters, and students also recalled more behaviors than carpenters when they assessed carpenters. At least for carpentry, these findings tend to support the notion that the recall of unfamiliar raters is largely behaviorally-based (Hauenstein & Kovach, 1988). However, carpenters' recall of carpentry should have been judgment-based. Unfortunately, this finding was not obtained when carpenters assessed the performance of other carpenters.

Although no direct support for the constant familiarity assumption was obtained in tests of these hypotheses, one interesting finding emerged from the analyses. Specifically, when recall was analyzed for each individual occupation, it was discovered that, whereas teaching performance led to judgment-based recall, carpentry performance led to behavior-based recall. These effects were seen regardless of subjects' levels of job familiarity.

One possible explanation as to why subjects' recalls were varied according to target occupation concerns the nature of the stimulus videotapes used in the present study. For example, the predominance of judgment-oriented recalls for the teaching occupation may have been accentuated by the nature of the stimulus videotapes. The lecture tapes used in the present study were not behaviorally-based, and performance level was varied by having each teacher exhibit a generally good, average, or poor lecture style. No distinct teaching behaviors (e.g., writing on the blackboard) were varied between performance conditions. Thus, teaching performance was not very conducive to behavioral recall. Conversely, the behaviorally-based recall of carpentry performance can be explained by the fact that the carpentry tapes used in the present study contained a number of discrete, critical incidents of ratee performance which may have been more amenable to behavioral rather than judgmental recall.

Results of the analyses conducted on subjects' seriation scores were not supportive of either of the two hypotheses. Again, the effects of the purpose manipulation probably accounted for the null findings obtained for hypothesis 1d. The results of the analyses conducted on subjects' seriation scores also provided no support for hypotheses 1e. That is, job familiarity had no effect of subjects' ordering of ratee performance information. One possible explanation for these unusual findings concerns the type of information-processing strategies used by familiar (i.e., expert) raters. Sujan (1985) found that, when faced with information discrepant with one's category knowledge, subjects familiar with a consumer product made more fine-grained distinctions (i.e., made more piecemeal-based evaluations) in their recalls than subjects who were unfamiliar with a product. Conversely, unfamiliar consumers made more category-based evaluations. Sujan claimed that piecemeal processes in evaluation should reflect statements of the specific operations performed on the information provided. This conclusion seems to be somewhat consistent with a seriation prediction. Although Sujan had her subjects rate objects and not people, Feldman (1988) has made a strong argument for the case that memory for persons involves the same cognitive processes as memory for objects.

Overall, results of the performance discrimination hypotheses were supported when familiar and unfamiliar raters assessed carpentry performance, but were not supported when familiar and unfamiliar raters assessed teaching performance. Although these latter findings are inconsistent with the hypotheses, the nature of the performances depicted in the teaching tapes may explain the why these results were obtained.

Examination of the means of the experts' ratings indicates that the teaching tapes contained a large degree of variability between each of the performance levels. Moreover, the high levels of convergent validities for both the behavioral and judgmental ratings of teachers shows that the experts raters were very much in agreement about the performances of the three ratees. Thus, the detection of performance differences between each of the three ratees may have been a simple task. However, examination of the means of the manipulation check items for teaching knowledge clearly shows that students were more familiar with teaching than were carpenters. This suggests that the manipulation of teaching performance was too powerful, and that differences between teaching levels were far greater than differences in job familiarity. Thus, the overpowering performance manipulation may have attenuated the effects of job familiarity on the discrimination of teaching performance. This explanation is also consistent with the effects of job familiarity on rating accuracy scores for the teaching occupation.

When carpentry was examined as the target occupation, results indicated that job familiarity had a direct impact on subjects' ratings of ratee performance, and this was evident on both behavioral and judgmental rating scales. For the behavioral ratings, while both groups of raters aptly discriminated between the average and poor carpenters, neither group distinguished between the performances of the average and good carpenters. Although the means were in the expected direction, the carpenters rated the good carpenter only slightly higher than the average carpenter. However, this finding needs to be qualified by the fact that the same outcome is seen when one examines the summed behavioral ratings supplied by the expert raters. Unlike the teaching tapes, examination of the pattern of convergent and discriminant validities for the expert carpenters indicates that these raters did not agree on the level of ratee performance when making their behavioral ratings. Thus, the carpenters in this study were actually correct in their failure to distinguish between average and good performance levels, at least on the behavioral rating scales. For judgmental ratings of carpentry performance, carpenters were incorrect in their failure to discriminate between average and good performance levels. Students, on the other hand, rated the average carpenter much higher than the good carpenter on both the behavioral and graphic rating scales. Thus, carpenters clearly discriminated between carpentry performance better than students.

One interesting finding seen in both sets of ratings made for the carpentry occupation is that students overrated the performance of the average carpenter. This is important because, in an effort to avoid systematic contrast effects (Murphy, Balzer, Lockhart, & Eisenman, 1985; Murphy, Gannett, Herr, & Chen, 1986), average performance was always seen first. Given these findings, unfamiliar raters may not have had an anchor by which to base their ratings against (Huber, Neale, & Northcraft, 1987). Huber et al. (1987) have argued that anchoring and adjustment (Kahneman & Tversky, 1973) may markedly influence performance ratings via raters' knowledge of appropriate performance standards. When assessing an occupation with which one is completely unfamiliar, the lack of any performance standards will have their greatest impact upon the initial performance. Subsequent performances (in this instance good and poor performances) should aid from the initial performance in that, after exposure to average performance, unfamiliar raters should have obtained some reference point upon which to base future ratings.

A second reliable finding obtained for both target occupations was that unfamiliar raters were consistently more lenient on both behavioral and judgmental ratings than familiar raters. These findings have been reported in previous literature (Barr & Hitt, 1986; Hakel, Ohnesorge, & Dunnette, 1970; Dipboye, Fromkin, & Wiback, 1975). For example, Barr and Hitt (1986) found leniency effects for unfamiliar raters (students) when providing favorability ratings and recommended starting salaries to professional interviewers. Additionally, job familiarity may influence the level of ratings given to ratees such that when raters are not entirely sure about the occupation with which they are rating, they tend to err on the conservative side.

Results of the rating accuracy measures are almost identical to the measures of performance discrimination. Familiar raters were more accurate than unfamiliar raters when assessing the carpentry occupation, and only when making judgmental ratings. In addition, the failure of the purpose manipulation probably led to the null findings obtained for the predicted interactions of appraisal purpose and job familiarity on accuracy of behavioral and judgmental ratings.

With respect to the hypothesized effects of job familiarity on rating accuracy, results of the accuracy scores for the teaching occupation showed only a marginally significant effect for job familiarity on differential accuracy of subjects' behavioral ratings. Although not statistically significant, students were more accurate at discriminating among teachers within behavioral dimensions than carpenters. However, since 7 out of the 8 accuracy measures were nonsignificant, and the only finding of importance was marginally significant, it is concluded that job familiarity had no effect on the accuracy of ratings made for teaching performance. As previously suggested, the powerful manipulation of performance level in the teaching videotapes probably contributed to the null findings obtained for job familiarity on the ratings of teaching performance.

Results of the accuracy scores for the carpentry occupation were mixed. Given the findings seen on the performance discrimination measures, it was expected that job familiarity would significantly influence accuracy scores on both behavioral and judgmental ratings. Unfortunately, job familiarity was seen to have an effect only on subjects' judgmental ratings of carpentry performance.

The null findings obtained for job familiarity on the behavioral ratings of carpentry are quite surprising. Although examination of the mean accuracy scores indicates some severity on behalf of the carpenters (i.e., lower elevation scores for the students), the means of the differential elevation and differential accuracy scores are not consistent with the raters' level ratings. Given that the students incorrectly rank-ordered the average and good carpenters, and the carpenters correctly rank-ordered the three carpenters, one would expect much better differential elevation and differential accuracy scores for the carpenters. This was not the case.

Perhaps one explanation why job familiarity had no effect on the carpentry behavioral ratings was that unfamiliar raters could merely record their observations of ratee performance without having to render any judgmental decisions. That is, perhaps no degree of expertise or familiarity is required for an individual to indicate how often (as opposed to how well) a person committed or failed to commit a certain behavior. Further, since unfamiliar raters would not likely have any preconceived dimensional schemata (Borman, 1978) their ratings merely reflected the actual behaviors that were observed. Therefore, since judgments were not involved for behavioral ratings, unfamiliar raters could simply rely on their memories for concrete details.

The results of the judgmental ratings made by subjects for the carpentry occupation are much more encouraging. Specifically, carpenters were more accurate than students on measures of elevation, differential elevation, and differential accuracy. Thus, when rating carpenters, except for stereotype accuracy, familiar raters provided more accurate judgmental ratings than unfamiliar raters. These findings are consistent with previous literature (Hauenstein & Walker, 1989; Kozlowski et al., 1986).

Two explanations are offered in an effort to explain why familiar raters were more accurate on judgmental carpentry ratings but not on the behavioral ratings. First, the choice of target occupations and rater populations in the current study may have produced more discrepant levels of job familiarity for carpentry than for teaching. In turn, this large difference in familiarity levels for carpentry may have resulted in carpenters being better able to translate their behavioral ratings into accurate judgmental ratings. Unfamiliar raters may not have had the knowledge structures to assimilate the behavioral information they had acquired into effective performance judgments (Chiesi et al., 1979). Thus, familiar raters may have an advantage over unfamiliar raters at rendering performance _judgments_ as opposed to _behaviors_.

Second, rating format differences may have contributed to the mixed findings obtained for job familiarity on carpentry ratings. Essentially, the behavioral rating scale employed in this study was a Behavioral Observation Scale (BOS; Latham & Wexley, 1977), whereas the judgmental rating scale was a simple graphic rating scale. More specifically, differences in the anchors used between the two rating formats may have influenced subjects' carpentry ratings. In general, the performance appraisal literature has failed to generate any consistent support for the BOS over other ratings formats for discriminating among ratees. Kane and Bernardin (1982) argue that this is because the anchors used on a typical BOS, whether they are intervals of percentages or verbal descriptions ranging from "none of the time" to "all of the time", may have different meanings for different job behaviors. That is, the anchors do not connote a constant level of performance satisfactoriness for all job behaviors. If this is true, then especially for carpentry, where discrete behaviors were rated along a continuum, subjects' ratings may not have truly reflected their perceptions of behavioral frequency.

In summary, the results of the accuracy scores are somewhat disappointing. Results showed that, for the two types of ratings made for the two different occupations, familiar raters were more accurate than unfamiliar raters only when making judgmental ratings for carpenters. Thus, in three out of four instances, the constant category assumption received support over the constant familiarity assumption. However, if conclusions are restricted to the carpentry occupation, then some support for the constant familiarity assumption is evident.

Discussion

To review, laboratory studies of performance appraisal imply one of two assumptions of processing invariance. First, the constant category assumption is implied in studies having undergraduate students rate unfamiliar occupations, and then generalizations are made to more familiar supervisors in real-world settings. Second, the constant familiarity assumption is implied in studies in which generalizations of the rating process are made only when raters rate an occupation with which they are familiar.

The present study attempted to examine these assumptions of processing invariance by manipulating two factors inherent in the performance appraisal process: job familiarity and appraisal purpose. Overall, it was predicted that, for different appraisal purposes, familiar raters would process performance information differently than unfamiliar raters. Appraisal purpose was employed as a means to indicate how differences in job familiarity lead to differences in the processes and outcomes of performance ratings. Therefore, differences in job familiarity were expected to provide support for the constant familiarity assumption over the constant category assumption. Several hypotheses were then generated which made specific predictions concerning the effects of job familiarity and appraisal purpose on the recall and ratings of ratee performance. This section will focus on the possible explanations and implications for the results obtained for these hypotheses.

Manipulation Effectiveness

Unfortunately, the manipulation of appraisal purpose was unsuccessful in the present study. This was unexpected because the manipulation of appraisal purpose (i.e., observational goals) has proven to be a robust phenomenon. For example, in the majority of studies that have varied subjects' observational goals, results have consistently shown increased memory for ratee information when subjects received an impression-set than when subjects received a memory-set. (cf. Hamilton, 1981; Hamilton, Katz, & Leirer, 1980a, 1980b; Srull, 1981, 1983; Wyer & Gordon, 1982). Although none of these studies employed manipulation check items, Foti and Lord (1987) used the same manipulation check items as in the present study, and found significant main effects for observational purpose on the extent to which subjects' formed impressions or memorized details of ratee performance. Thus, in previous studies, observational purpose has been shown to be an effective manipulation.

Two explanations can be offered as to why appraisal purpose was ineffective in the present experiment. First, the use of target occupation as a repeated measure may have weakened the effects of appraisal purpose over time. However, as reported above, analyses conducted on subjects' responses from trial 1 only indicated this not to be the case. Second, in almost all social cognition studies that have varied observational purpose, the instructions given to subjects about their processing objectives were much more detailed than in the present study (Markus et al., 1985; Hoffman, Mischel, & Mazze, 1981; Hamilton et al., 1980a, 1980b). The important difference between these manipulations and the manipulation of purpose in the present study is that subjects were not told _why_ they had to remember as much as possible, or why they had to form an impression of the ratees' performances. That is, subjects were not informed that they would be asked to later recall ratee information (memory-set) or make judgments about ratee performance (impression-set). This type of manipulation was avoided in the present study because typical performance appraisal studies do not include explicit notations about the type of information required of subjects. Therefore, the lack of a stronger "set" given to subjects prior to observation of ratee performance, including the type of ratings to be made, may have weakened the manipulation effectiveness of appraisal purpose.

The failure of the purpose manipulation in the present study has implications for the tests of processing invariance. Specifically, appraisal purpose was included in the design in order to demonstrate how job familiarity influences the processing of performance information. In turn, differences in the rating process between familiar and unfamiliar raters would indicate support for the constant familiarity assumption over the constant category assumption. However, since the purpose manipulation was ineffective, the test of the assumptions of processing invariance was limited in the present study. For example, any hypotheses predicting interactions between appraisal purpose and job familiarity must now be interpreted with caution, since an adequate manipulation of purpose was not seen in the present investigation. What this implies is that job familiarity alone must suffice as the vehicle for testing the assumptions of processing invariance.

Assumptions of Processing Invariance

In summary, the results of the hypotheses pertaining to the process measures generally failed to support the constant familiarity assumption. No major differences were seen in the processing of ratee performance information according to subjects' levels of job familiarity. Unfamiliar raters did not recall significantly more behaviors than familiar raters, and familiar raters did not recall significantly more judgments than unfamiliar raters. Results indicated, however, that the recalls made by raters varied systematically according to the occupation being rated. For example, across both rater populations, subjects recalled more judgments when rating teachers and more behaviors when rating carpenters. For the outcome measures, the hypotheses concerning effects for job familiarity were supported only when raters rated carpenters, and these effects were primarily seen when subjects provided judgmental ratings of carpentry performance. For example, while the constant familiarity assumption was supported for both behavioral and judgmental ratings of carpentry performance discrimination, differences in job familiarity on measures of rating accuracy emerged only for subjects' judgmental carpentry ratings.

With regards to the processing of performance information, the absence of consistent effects for job familiarity on the free recall measures indicates that the _processes_ used to arrive at performance ratings may be similar across raters. That is, raters may search, store, organize, and recall ratee behavior similarly, regardless of their levels of job familiarity. These results are inconsistent with previous literature which has examined the cognitive processes of familiar and unfamiliar raters (e.g., Hauenstein & Kovach, 1988; Hauenstein & Walker, 1989; Kozlowski & Kirsch, 1987; Markus et al., 1985). However, conclusions drawn from the current study must be tempered by the failure of the purpose manipulation and nature of the stimulus materials employed.

More specifically, the teaching tapes used in the present study were based on overall teaching style, and were thus not behaviorally-based. On the other hand, the carpentry tapes were mainly composed of discrete, critical incidents of woodworking behaviors. Essentially then, the teaching tapes were more conducive to judgmental recall, whereas the carpentry tapes were more prone to behavioral recall. The use of such tapes in the present study may have thus limited the potency of the tests processing invariance. Since each occupation demanded a certain type of recall, job familiarity had little influence on the recall of performance information.

When one examines the results of the outcome measures, the effects of job familiarity become more apparent. For measures of performance discrimination, job familiarity had significant effects on both behavioral and judgmental carpentry ratings. For measures of rating accuracy, job familiarity only influenced judgmental ratings of carpentry performance. On the other hand, job familiarity had no effect on the discrimination or accuracy of ratings made for teaching performance. These findings thus provide partial support for both the constant familiarity and the constant category assumptions.

As previously suggested, the absence of effects for job familiarity on the outcome measures of teaching performance can be attributed to the simplicity of the rating task. The levels of performance depicted in the teaching tapes were too divergent for differences in job familiarity to have an effect on discrimination and accuracy of teaching behavior. Support for the constant category assumption is then tempered by the fact that differences in teaching familiarity were not allowed to be adequately tested in the present study.

Results of the outcome measures for carpentry performance indicate greater support for the constant familiarity assumption than for the constant category assumption. This is consistent with previous literature which has examined the ratings of familiar and unfamiliar raters (Hauenstein & Walker, 1989; Kozlowski & Kirsch, 1987; Smither & Reilly, 1989). Given the nature of the stimulus materials used in the present study, these findings should be given more credence than the findings obtained for teaching performance. First, since the carpentry tapes contained less divergent levels of ratee performance, true differences in carpentry aptitude were more accurately reflected in these videotapes than were differences in teaching aptitude reflected in the teaching tapes. Further, these more subtle performance differences are more likely to be seen in real-world performance appraisal settings. Second, the choice of carpentry as a target occupation is also likely to be independent of subjects' knowledge of other occupations (i.e., teaching). That is, the manipulation of job familiarity in the present study may have been unbalanced such that carpenters were less unfamiliar with teaching than students were unfamiliar with carpentry. Thus, the results of the ratings made by familiar and unfamiliar raters for the assessment of carpentry performance may be more indicative of the true effects of job familiarity on performance ratings.

In sum, results obtained for ratings of carpentry performance offer some optimism for the constant familiarity assumption. Based on these results, future studies which examine occupations not related to the academic setting should continue to demonstrate the differential effects of job familiarity on performance ratings. Moreover, these findings suggest that studies which have undergraduates rate occupations outside of the academic setting with which they are unfamiliar may not be generalizable to real-world settings where supervisors rate jobs with which they are much more familiar.

For both the process and outcomes of performance appraisal seen in the current study, the findings discussed above inevitably point to the _content_ of the stimuli presented to subjects. Results of the process measures indicated different types of recall for different occupations, regardless of job familiarity. For the outcome measures, job familiarity was only a factor on subjects' ratings when they rated carpentry and not when they rated teaching. Together, these findings suggest that the content of the stimulus materials presented to subjects in the current study may have been a major factor in the recall and rating of ratee performance. That is, the results of the present study may have been _stimulus-bound_ (cf. Funder, 1987).

For example, results of the process measures show that across raters, recall of teaching performance was primarily judgment-based and recall of carpentry performance was behavior-based. These findings have implications for laboratory research on the performance appraisal process. Since all raters generally recalled more evaluative statements regarding teaching performance, other studies which have raters rate teaching style may also obtain judgment-oriented recalls. Yet, it should be noted that the teaching tapes used in the present study did not include discrete, critical incidents of lecture performance. Rather, raters rated the lecturers' _style_ of teaching. Laboratory studies using teachers as target ratees may thus be content-specific (Funder, 1987). Had the teaching tapes in the present study been developed to emphasize teaching _behaviors_, subjects' recalls may have been entirely different. What this suggests then is that studies which have used teaching as the stimulus domain (e.g., Murphy et al., 1982; Murphy et al., 1985; Murphy et al., 1986; Smither et al., 1988; Smither & Reilly, 1989; Krzystofiak, Cardy, & Newman, 1988) may also be stimulus-bound. That is, results obtained in laboratory studies of performance appraisal where students rate teachers may be bound by the style of lecturing seen on the tapes, and thus, any findings seen may not be generalizable to organizational settings.

The fact that the recall of carpentry performance was primarily behaviorally-based is also consistent with Funder's (1987) notion that social judgments formed in laboratory settings are content-dependent. Perhaps the concrete nature of the tasks (hammering, sawing, staining, and sanding) included in the carpentry tapes promoted memory for behavior rather than memory for judgment. This finding would then add more evidence to the argument that laboratory studies of the rating process are bound by the content of the stimuli presented to raters. Similarly, other performance appraisal studies concerned with the rating process may have obtained results consistent with their social stimuli rather than with the content of the performance domain presented to raters.

The content of the stimuli used in the present study may also have accounted for the findings obtained for job familiarity on subjects' performance ratings. In short, job familiarity only influenced subjects' ratings of carpenters. Again, the content of the teaching tapes led to null effects for job familiarity, whereas the the carpentry tapes allowed for differences in job familiarity to be seen in subjects' performance discrimination measures and rating accuracy scores. Thus, it seems apparent that a better test of the effects of job familiarity, and the assumptions of processing invariance, would involve manipulations of several different occupational categories in the same study.

The conclusion that results of laboratory studies of performance appraisal may be stimulus-bound is consistent with the arguments made by Banks and Murphy (1985) and Ilgen and Favero (1985). These authors have listed a number of contextual factors that have widened the gap between appraisal research and practice, including observation of ratee performance over time and the consequences of performance ratings. Although the present study sought to examine a different issue related to the generalizability of performance appraisal research (job familiarity), future research can examine the issue of content-specific stimuli empirically by either: (a) varying the stimuli-content of one occupational category presented to raters; or (b) holding constant the stimulus-content of two or more occupational categories presented to raters. Another possible solution to this problem would be to abandon the use of "staged" videotaped performances altogether in laboratory research on performance appraisal. Hauenstein and Walker (1989) used actual footage of a football game--complete with true scores provided by college coaches and actual performance discrepancies within and between players--in an effort to more accurately portray a real-world appraisal setting. Stimulus materials such as these are advocated for future appraisal research because they avoid the artificiality of staged videotaped performances, and thus, are better able to overcome the problems of stimulus-bound material. When such research is established, more definitive conclusions about can be made about the generalizability of laboratory studies of the performance appraisal process.

Implications

Although the results of this study were largely nonsupportive of the hypotheses, implications to both research and application can be derived. It is these implications that are the focus of this section. First, the use of social cognition measures (e.g., free recall) is strongly encouraged for laboratory studies of the cognitive processes of performance appraisal. The study of how raters process performance information is currently a widely investigated topic in the industrial and organizational psychology literature. However, many studies that claim to examine raters' cognitive processes fail to incorporate process measures. Unless such measures are incorporated regularly into the design of the study, any conclusions dealing with raters' use of schemas, prototypes, or categories can only be indirectly inferred from subjects' performance ratings.

A second implication concerns the major focus of this study, namely, the generalizability of laboratory studies of performance appraisal. Banks and Murphy (1985) and Ilgen and Favero (1985) have pointed out a number of contextual factors that are absent in laboratory performance appraisal studies which limit their generalizability to applied settings. One of these variables is appraisal purpose, a variable germane to most appraisal contexts. Although the manipulation of purpose was unsuccessful in the current study, future research could examine its role by crossing different types of purpose (i.e., administrative versus feedback) with other relevant organizational variables.

The current study chose to focus on two different, but related issues of generalizability: whether the persons selected as raters influence performance ratings, and whether the occupation to be rated influences performance ratings. These two factors, then, comprised the manipulation of job familiarity. Too often, laboratory studies of performance appraisal rely on undergraduate sophomores as their sole source of performance ratings. While this choice of rater population may be adequate for some appraisal studies, it may not be the best choice for studies that wish to make generalizations to applied settings. In addition, the use of raters who are beyond the academic setting provides an opportunity for researchers to examine an important potential moderator of many observed relationships seen in the appraisal literature. Moreover, many researchers assume that the findings obtained using undergraduates as subjects tend to be conservative in nature. One way of empirically testing this notion is to include levels of raters as a factor in future research studies.

A third implication of this research is the potential impact of individual difference variables on the rating process. By and large, individual difference variables have been ignored in the appraisal literature. In the present study, job familiarity served two purposes. First, job familiarity represented a vehicle for testing the two competing assumptions of processing invariance. Second, job familiarity represented an important individual difference variable which may also moderate many relations between contextual variables and performance ratings. Future research should continue to examine the role of job familiarity, as well as other individual difference variables in the performance appraisal context.

This study also has implications for applied settings. On an intuitive level, job familiarity would appear to be an important variable in the rendering of performance appraisals in organizational settings. Supervisors who are highly familiar with the job with which they are rating should seemingly provide, if not more accurate ratings, better feedback to their subordinates. Moreover, subordinates will likely be more accepting of feedback when they perceive it as coming from a supervisor who is familiar with a job (e.g., Stone, Guetal, & McIntosh, 1984). On an empirical level, job familiarity did result in leniency effects, with unfamiliar raters being more lenient than familiar raters. It is thus recommended that organizations ensure that their raters have either direct experience with the job they are rating, or have been well trained in the details of the nature of their subordinates' work.

Limitations

This study had several limitations. First, the manipulation of appraisal purpose was unsuccessful. The levels of purpose used in this study (memory-set vs. impression-set) were chosen because it was felt that, in conjunction with job familiarity, they would provide the best test of the assumptions of processing invariance. Future studies could vary different, and perhaps more applied levels of appraisal purpose (see above). Second, the manipulation of job familiarity in the present study was apparently unbalanced. This is because students were more unfamiliar with carpentry than carpenters were unfamiliar with teaching. With this in mind, it is felt that future research which aptly manipulates job familiarity according to choice target occupations and rater populations will likely provide a better test of the assumptions of processing invariance. Third, given that some raters were expected to discriminate between ratee performance levels and some were not, and that subjects' recalls were expected to differ depending on their levels of job familiarity, the teaching tapes used in the present study were perhaps inappropriate. Job familiarity may have had stronger effects in both the process and outcome measures had the teaching tapes contained distinct critical incidents of teaching performance, and more subtle differences between performance levels. Results obtained when using teaching videotapes may thus be bound to the laboratory context in which they were presented. Finally, one last limitation concerns the setting of the study. This study was conducted in a laboratory setting without any of the usual distractions seen in the typical appraisal context. Future studies which vary the context in which appraisals are made will be more generalizable to nonacademic settings.

Despite the above limitations, this study had several strong points. First, all raters rated multiple ratees at the same time. This situation is seen as more closely approximating the complex appraisal situation seen outside the usual laboratory setting. Second, a rater population outside of the academic setting was employed in this study. Thus, the external validity of this study was not constrained by the sole use of undergraduates. Lastly, rating process as well as rating accuracy were measured directly, using measures of free recall and distance accuracy, respectively. Future research examining the processes and outcomes of performance appraisal should continue to incorporate such measures in their design.

Conclusions

To summarize, the purpose of this study was to shed light on the factors which may moderate the extent to which laboratory studies of the cognitive processes of performance appraisal are generalizable across raters. Essentially, job familiarity and appraisal purpose were manipulated in order to provide a test of the two assumptions of processing invariance: the constant category assumption and the constant familiarity assumption. The constant category assumption holds that regardless of the occupation being rated, all raters will basically use the same processes to arrive at performance ratings; thus, ratings will not differ as a function of familiarity with the occupation being rated. The constant familiarity assumption makes the opposite prediction, namely, that the processes and outcomes of performance ratings will differ as a function of the raters' levels of job familiarity. Support for the constant category assumption would imply that research conducted using undergraduates as raters rating unfamiliar occupations is indeed generalizable to all raters. Support for the constant familiarity assumption would imply that only those studies in which raters rated familiar occupations (e.g., experiments having students rate teachers) are generalizable to all raters.

The results of this study provided greater support for the constant category assumption than for the constant familiarity assumption. Apparently, subjects' levels of job familiarity did not have a great impact on the recall or ratings of performance information. However, several unique aspects of the study's design, along with some findings supporting the constant familiarity assumption, preclude the recommendation that research in the appraisal domain continue to use undergraduates rating unfamiliar occupations.

First, the stimulus materials used for the teaching occupation may have attenuated the effects of job familiarity on subjects' recall and ratings of ratee performance information. Second, when focusing attention solely on the carpentry occupation, the results provided mixed support for both assumptions of processing invariance. Given the results of the accuracy measures for both behavioral and judgmental carpentry ratings, questions can be raised concerning the generalizability of performance appraisal studies which rely solely on undergraduates as raters. This should be of particular concern when appraisal studies are conducted in which students are rating unfamiliar occupations. In such cases, caution is suggested when generalizations are made to real-world supervisors, because these raters are apt to much more familiar with the occupations they are rating.

References

Athey, T.R. & McIntyre, R.M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory. Journal of Applied Psychology, 72, 567-572.

Balzer, W.K. (1986). Biases in the recording of performance-related information: The effects of initial impression and centrality of the appraisal task. Organizational Behavior and Human Decision Processes, 37, 329-347.

Banks, c.G. & Murphy, K.R. (1985). Toward narrowing the research- practice gap in performance appraisal. Personal Psychology, 38, 335-345.

Bargh, J.A. (1982). Attention and automaticity in the processing of self-relevant information. Journal of Personality and Social Psychology, 43, 425-436.

Bargh, J.A. & Thein, R.D. (1985). Individual construct accessibility, person memory, and the recall-judgment link: The case of information overload. Journal of Personality and Social Psychology, 49, 1129-1146.

Barr, S.H. & Hitt, M.A. (1986). A comparison of selection decision models in manager versus student samples. Personnel Psychology, 39, 599-617.

Becker, B.E. & Cardy, R.L. (1986). Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. Journal of Applied Psychology, 71, 662-671.

Bernardin, H.J. & Beatty, R.W. (1984). Performance Appraisal: Assessing Human Behavior at Work. Boston, MA: Kent.

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

Bernstein, V., Hakel, M.D., & Harlan, A. (1975). The college student as interviewer: A threat to generalizability? Journal of Applied Psychology, 60, 266-268.

Binning, J.F., Zaba, A.J., & Whatham, J.C. (1986). Explaining the biasing effects of performance cues in terms of cognitive categorization. Academy of Management Journal, 29, 521-535.

Bobrow, D.G. & Norman, D. (1975). Some principles of memory schemata. In D.G. Bobrow & A. Collins (Eds.), Representation and understanding: Studies in cognitive science. New York: Academic Press.

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

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. (1983). Implications of personality theory and research for the rating of work performance in organizations. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance Measurement and Theory.* Hillsdale, N.J.: Erlbaum

Cafferty, T.P., DeNisi, A.S., & Williams, K.J. (1986). Search and retrieval patterns for performance information: Effects on evaluations of multiple targets. *Journal of Personality and Social Psychology*, *50*, 676-683.

Cantor, N., Mischel, W. & Schwartz, J. (1982). Categorical knowledge about the social world: Structure, content, use and abuse. In A. Hastorf & A. Isen (Eds.), *Cognitive Social Psychology.* New York: Elsevier North-Holland.

Chiesi, H.L., Spilich, G.J., & Voss, J.F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18,* 257-273.

Christensen-Szalansi, J.J.J. & Beach, L.R. (1984). The citation bias: Fad and fashion in the judgment and decision literature. *American Psychologist, 39,* 75-78.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* (rev. ed.). New York: Academic Press.

Cohen, C. & Ebbesen, E.B. (1979). Observational goals and schema activation: A theoretical framework for behavior perception. *Journal of Experimental Social Psychology, 15,* 305-329.

Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Boston, MA: Houghton-Mifflin.

Cooper, W.H. (1981). Ubiquitous halo. *Psychological Bulletin, 90,* 218-244.

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

DeNisi, A.S., Cafferty, T.P., & Meglino, B.M. (1984). A cognitive model of the performance appraisal process: A model and research propositions. Organizational Behavior and Human Performance, 33, 360-396.

DeNisi, A.S., Cornelius, E.T., & Blencoe, A.G. (1987). Further investigation of common knowledge effects on job analysis ratings. Journal of Applied Psychology, 72, 262-268.

Dipboye, R.L., Fromkin, H.L., & Wiback, K. (1975). Relative importance of applicant sex, attractiveness, and scholastic standing in evaluation of job applicant resumes. Journal of Applied Psychology, 60, 39-43.

Ebbesen, E.B. & Konecni, V.J. (1980). On the external validity of decision-making research: What do we know about decisions in the real world? In T.S. Wallsten (Ed.), Cognitive Processes in Choice and Decision Behavior. Hillsdale, N.J.: Erlbaum.

Feldman, J. (1988). Objects in categories and objects as categories. In T.K. Srull & R.S. Wyer (Eds.), Advances in Social Cognition: A Dual Process Model of Impression Formation (Vol. 1 pp. 53-62). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 66, 127-148.

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

Fiske, S.T. & Linville, P.W. (1980). What does the schema concept buy us? Personality and Social Psychology Bulletin, 6, 543-557.

Fiske, S.T. & Taylor, S.E. (1984). Social Cognition. New York, N.Y.: Random House.

Foti, R.J. & Lord, R.G. (1987). Prototypes and scripts: The effects of alternative methods of processing information on rating accuracy. Organizational Behavior and Human Decision Processes, 39, 318-340.

Friedman, L. & Harvey, R.J. (1986). Can raters with reduced job descriptive information provide accurate position analysis (PAQ) ratings? Personnel Psychology, 39, 779-790.

Funder, D.C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. Psychological Bulletin, 101, 75-90.

Gordon, M.E., Slade, L.A., Schmitt, H. (1986). The "science of the sophomore" revisited: From conjecture to empiricism. Academy of Management Review, 11, 191-207.

Greenberg, J. (1987). The college sophomore as guinnea pig: Setting the record straight. Academy of Management Review, 12, 157-159.

Hakel, M.D., Ohnesorge, J.P., & Dunnette, M.D. (1970). Interviewer evaluations of job applicants' resumes as a function of the qualifications of the immediately preceding applicants: An examination of contrast effects. Journal of Applied Psychology, 54, 27-30.

Hamilton, D.L. (1981). Cognitive representations of persons. In E. Higgins, C. Herman, & M. Zanna (Eds.), Social Cognition: The Ontario Symposium, Vol. 1. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Hamilton, D.L., Katz, L.B., & Leirer, V.O. (1980a). Cognitive representation of personality impression: Organizational processes in first impression formation. Journal of Personality and Social Psychology, 39, 1050-1063.

Hamilton, D.L., Katz, L.B., & Leirer, V.O. (1980b). Organizational processes in impression formation. In R. Hastie, T.M. Ostrom, E.B. Ebbesen, R.S. Wyer, D.L. Hamilton, & D.E. Carlston (Eds.), Person Memory: The Cognitive Basis of Social Perception. Hillsdale, N.J.: Erlbaum.

Harari, O. & Zedeck, S. (1973). Development of behaviorally anchored rating scales for the evaluation of faculty teaching. Journal of Applied Psychology, 58, 261-265.

Hartwick, J. (1979). Memory for trait information: A signal detection analysis. Journal of Experimental Social Psychology, 15, 533-552.

Hastie, R. (1980). Memory for behavioral information that confirms or contradicts a personality impression. In R. Hastie, T.M. Ostrom, E.B. Ebbesen, R.S. Wyer, D.L. Hamilton, & D.E. Carlston (Eds.), Person Memory: The Cognitive Basis of Social Perception. Hillsdale, N.J.: Erlbaum.

Hastie, R. (1981). Schematic principles in human memory. In E.T. Higgins, C.P. Herman, & M.P. Zanna (Eds.), Social Cognition: The Ontario Symposium. (Vol. 1). Hillsdale, N.J.: Erlbaum.

Hastie, R. & Kumar, P.A. (1979). Person Memory: Personality traits as organizing principles in memory for behaviors. Journal of Personality and Social Psychology, 37, 25-38.

Hastie, R. & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task in memory-based or on-line. Psychological Review, 93, 258-268.

Hastie, R. , Park, B, & Weber, R. (1984). Social Memory. In R.S. Wyer & T.K. Srull (Eds.), Handbook of Social Cognition , (Vol. 2. pp. 151-211). Hillsdale, N.J.: LEA.

Hauenstein, N.M.A. & Foti, R.J. (1987). Effects of increasing information processing demands on the rating process. Unpublished Manuscript. 115-169.

Hauenstein, N.M.A. & Kovach, R.C. (1988). The effects of category familiarity on the rating process Paper presented at the 96th annual Convention of the American Psychological Association, Atlanta, GA.

Hauenstein, N.M.A. & Walker, S.E. The relationship between memory and appraisal accuracy: The moderating effects of job knowledge and judgment complexity. Paper presented at the annual convention of the Society for Industrial and Organizational Psychology, April, 1989. Boston, MA.

Higgins, E.T., King, G., & Mavin, G.H. (1982). Individual construct accessibility and subjective impressions and recall. Journal of Personality of Social Psychology, 43, 35-47.

Hoffman, C., Mischel, W., & Mazze, K. (1981). The role of purpose in the organization of information about behavior: Trait-based versus goal-based categories in person perception. Journal of Personality and Social Psychology, 40, 211-225.

Hogarth, R.M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. Psychological Bulletin, 90, 197-217.

Huber, V.L., Neale, M.A., & Northcraft, G.B. (1987). Judgment by heuristics: Effects of ratee and rater characteristics and performance standards on performance-related judgments. Organizational Behavior and Human Decision Processes, 40, 149-169.

Hunt, E., Frost, N., & Lunneborg, C. (1973). Individual differences in cognition: A new approach to intelligence. In G.H. Bower (Ed.), The Psychology of Learning and Motivation. New York: Academic Press.

Ilgen, D.R. & Favero, J.L. (1985). Limits in the generalization from psychological research to performance appraisal processes. Academy of Management Review, 10, 311-321.

Johnson, P.E., Duran, A.S., Hassebrock, F., Moller, J., Prietula, M., Feltovick, P.J., & Swanson, D.B. (1981). Expertise and error in diagnostic reasoning. Cognitive Science, 5, 235-283.

Kane, J.S., & Bernardin, H.J. (1982). Behavioral observation scales and the evaluation of performance appraisal effectiveness. Personnel Psychology, 35, 635-641.

Kavanagh, M.J., MacKinney, A.C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analysis of ratings. *Psychological Bulletin, 75,* 34-49.

Keppel, G. (1982). *Design and Analysis: A Researcher's Handbook.* New Jersey: Prentice-Hall.

Kelly, G.A. (1955). *The Psychology of Personal Constructs.* New York: Norton.

Kelley, H.H. (1973). The processes of causal attribution. *American Psychologist, 28,* 107-128.

Kingström, P.O. & Mainstone, L.E. (1985). An investigation of the rater-ratee acquaintance and rater bias. *Academy of Management Journal*, *28*, 641-653.

Kozlowski, S.W.J. & Ford, J.J. (1988). Effects of familiarity, performance, constraint and memory on rater information acquisition strategies. In A.S. DeNisi (Chair), *Memory issues in the performance appraisal process*. Symposium presented at the Annual Meeting of the Academy of Management, Anaheim, CA.

Kozlowski, S.W.J. & Kirsch, M.P. (1987). The systematic distortion hypothesis, halo, and accuracy: An individual level analysis. *Journal of Applied Psychology*, *72*, 252-261.

Kozlowski, S.W.J., Kirsch, M.P., & Chao, G.T. (1986). Job knowledge, ratee familiarity, conceptual similarity, and halo error: An exploration. *Journal of Applied Psychology*, *71*, 45-49.

Krzystofiak, F, Cardy, R., & Newman, J. (1988). Implicit personality and performance appraisal: The influences of trait inferences on on evaluations of behavior. *Journal of Applied Psychology*, *73*, 515-521.

Landman, J. & Manis, M. (1983). Social cognition: Information processing and social theoretical perspectives. In L. Berkowitz (Ed.), *Handbook of Social Cognition* (Vol. 16 pp. 49-123). Hillsdale, N.J.: LEA.

Landy, F.J. (1987). Presentation in K.Williams (Chair), *Cognitive Research in I/O Psychology: Challenges for the future.* Symposium conducted at the meeting for the Society for Industrial and Organizational Psychology, Atlanta, GA.

Landy, F.J. & Bates, F. (1973). Another look at contrast effects in the employment interview. *Journal of Applied Psychology, 58,* 141-144.

Landy, F.J. & Farr, J.L. (1980). Performance rating. *Psychological Bulletin*, *87*, 72-107.

Landy, F.J. & Farr, J.L. (1983). The Measurement of Work Performance. New York: Academic Press.

Latham, G.P., & Wexley, K.N. (1977). Behavioral observation scales for performance appraisal. Personnel Psychology, 30, 255-268.

Lichtenstein, M. & Srull, T.K. (1985). Conceptual and methodological issues in examining the relationship between consumer memory and judgment. In L.F. Alwitt & A.A. Mitchell (Eds.), Psychological Processes in Advertising Effects: Theory, Research, and Application. (pp. 113-128). Hillsdale, N.J.: Erlbaum.

Lingle, J.H. & Ostrom, T.M. (1979). Retrieval selectivity in memory-based impression judgments. Journal of Personality of Social Psychology, 37, 1098-2109.

Locke, E.A. (1986). Generalizing from laboratory to field: Ecological validity or abstraction of essential elements. In E.A. Locke (Ed.), Generalizing from laboratory to field settings. (pp. 3-9). Lexington, M.A.: Lexington Books.

Lynch, J.G., Jr. (1983). The role of external validity in theoretical research. Journal of Consumer Research, 10, 109-111.

Markus, H. (1977). Self-schemata and processing information about the self. Journal of Personality of Social Psychology, 35, 63-78.

Markus, R.W., Smith, J., & Moreland, R.L. (1985). Role of the self-concept in the perception of others. Journal of Personality and Social Psychology, 49, 1494-1512.

McCormick, E.J., Jeanneret, P.R., & Mecham, R.C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 56, 347-368.

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.

Mount, M.K. & Thompson, D.E. (1987). Cognitive categorization and quality of performance ratings. Journal of Applied Psychology, 72, 240-246.

Murphy, K.R. (1988). Personal communication.

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

Murphy, K.R., Balzer, W.K., Lockhart, M.C., Eisenman, E.J. (1985). Effects of previous performance on evaluations of present performance. Journal of Applied Psychology, 70, 72-84.

Murphy, K.R., Gannett, B.A., Herr, B.M., & Chen, J.A. (1986). Effects of subsequent performance on evaluations of present performance. Journal of Applied Psychology, 71, 427-431.

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.

Murphy, M.D. & Puff, C.R. (1982). Free recall: Basic methodology and analyses. In C.R. Puff (Ed.), Handbook of research methods in human memory and cognition (pp. 99-128). New York: Academic Press.

Nathan, B., & Lord, R. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. Journal of Applied Psychology, 68, 102-114.

Neale, M.A. & Northcraft, G.B. (1986). Experts, amateurs, and refrigerators: Comparing expert and amateur negotiators in a novel task. Organizational Behavior and Human Decision Processes, 38, 305-317.

Pelligrino, J.W. & Battig, W.F. (1974). Relationships among higher order organizational measures and free recall. Journal of Experimental Psychology, 102, 463-472.

Peters, L.H. & DeNisi, A.S. (1988). An information processing role for appraisal purpose and job type in the development of appraisal systems. Paper presented at the 96th Annual Convention of the American Psychological Association, Atlanta, GA.

Phillips, J.S. & Lord, R.G. (1982). Schematic information processing and perceptions of leadership in problem-solving groups. Journal of Applied Psychology, 67, 486-492.

Pulakos, E. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

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

Roenker, D.L., Thompson, C.P., & Brown, S.C. (1971). Comparison of meaures for the estimation of clustering in free recall. Psychological Bulletin, 76, 45-48.

Rosenthal, R. & Rosnow, R.L. (1975). The Volunteer Subject. New York: Wiley-Interscience.

Rosenthal, R. & Rosnow, R.L. (1984). Essentials of behavioral research: Methods and data analysis. New York: McGraw-Hill.

Rothbart, M., Fulero, S., Jensen, C., Howard, J., & Birrell, P. (1978). From individual to group impressions: Availability heuristics in stereotype formation. Journal of Experimental Social Psychology, 14, 237-255.

Rothbart, M., Evans, & Fulero, S. (1979). Recall for confirming events: Memory processes and the maintenance of social stereotypes. Journal of Experimental Social Psychology, 15, 343-355.

Rumelhart, D.E. (1984). Schemata and the cognitive system. In R.S. Wyer and T.K. Srull (Eds.), Handbook of Social Cognition , (Vol. 2 pp. 161-188). Hillsdale, N.J.: LEA.

Sherman, S.J., Zener, & Johnson, J., & Hirt, E.R. (1983). Social explanation: The role of timing, set, and recall on subjective likelihood estimates. Journal of Personality and Social Psychology, 44, 1127-1143.

Smith, J.E. & Hakel, M.D. (1979). Convergence among data sources, response bias, and reliability and validity of a structured job analysis questionnaire. Personnel Psychology, 32, 677-692.

Smither, J.W. & Reilly, R.R. (1987). True intercorrelation among job components, time delay in rating, and rater intelligence as determinants of accuracy in performance ratings. Organizational Behavior and Human Decision Processes, 40, 369-391.

Smither, J.W. & Reilly, R.R. (1987). Relationship between job knowledge and the reliability of conceptual similarity schemata. Journal of Applied Psychology, 74, 530-534.

Smither, J.W., Reilly, R.R., & Buda, R. (1988). Effects of prior performance information on ratings of present performance: Contrast versus assimilation revisited. Journal of Applied Psychology, 73, 487-496.

Srull, T.K. (1981). Person Memory: Some tests of associative storage and retrieval models. Journal of Experimental Psychology: Human Learning and Memory, 7, 440-463.

Srull, T.K. (1983). Organizational and retrieval processes in person memory: An examination of processing objectives, presentation format, and the possible role of self-generated retrieval cues. Journal of Personality and Social Psychology, 44, 1157-1170.

Srull, T.K. (1984). Methodological techniques for the study of person memory and social cognition. In R.S. Wyer and T.K. Srull (Eds.), Handbook of Social Cognition , (Vol. 2 pp. 1-72). Hillsdale, N.J.: LEA.

Srull, T.K., Lichtenstein, M., & Rothbart, M. (1985). Associative storage and retrieval processes in person memory. Journal of Experimental Psychology, 11, 316-345.

Srull, T.K & Wyer, R.S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. Journal of Personality and Social Psychology, 31, 1660-1672.

Srull, T.K & Wyer, R.S. (1980). Category accessibility and social perception: Some implications for the study of person memory and interpersonal judgment. Journal of Personality and Social Psychology, 38, 841-856.

Srull, T.K. & Wyer, R.S. (1986). The role of chronic and temporary goals in social information processing. In R.M. Sorrentino & E.T. Higgins (Eds.), Handbook of Motivation and Cognition (pp.503-549). New York: Guilford.

Stone, D.L., Guetal, H.G., & McIntosh, B. (1984). The effects of feedback sequence and expertise on perceived feedback accuracy. Personnel Psychology, 37, 487-506.

Sujan, M. (1985). Consumer knowledge: Effects on evaluation strategies mediating consumer judgments. Journal of Consumer Research, 12, 31-46.

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.

Taylor, S.E. & Crocker, J. (1981). Schematic bases of social information processing. In E.T. Higgins, C.P. Herman, & M.P. Zanna (Eds.), Social Cognition: The Ontario Symposium. (Vol. 1., pp. 89-134). Hillsdale, N.J.: Erlbaum.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5, 207-232.

Williams, K.J. (1988). Personal communication.

Williams, K.J., DeNisi, A.S., Blencoe, A.G., & Cafferty, T.P. (1985). The role of appraisal purpose: Effects of purpose on information acquisition and utilization. Organizational Behavior and Human Decision Processes, 36, 314-339.

Williams, K.J., DeNisi, A.S., Meglino, B.M., & Cafferty, T.P. (1986). Initial decisions and subsequent performance ratings. Journal of Applied Psychology, 71, 189-195.

Wyer, R.S., Bodenhausen, G.V., & Srull, T.K. (1984). The cognitive representation of persons and groups and its effect on recall and recognition memory. *Journal of Experimental Social Psychology, 20,* 445-469.

Wyer, R.S. & Gordon, S.E. (1982). The recall of information about persons and groups. *Journal of Experimental Social Psychology, 18,* 128-164.

Wyer, R.S. & Gordon, S.E. (1984). The cognitive representation of social information. In R.S. Wyer and T.K. Srull (Eds.), *Handbook of Social Cognition,* (Vol. 2. pp. 73-150). Hillsdale, N.J.: Erlbaum

Wyer, R.S., & Srull, T.K. (1986). The processing of social stimulus information: A conceptual integration. In R. Hastie, T.M. Ostrom, E.B. Ebbesen, R.S. Wyer, D.L. Hamilton, & D.E. Carlston (Eds.), *Person Memory: The Cognitive Basis of Social Perception* (pp. 227-300). Hillsdale, N.J.: Erlbaum.

Wyer, R.S., Srull, T.K., & Gordon, S.E. (1984). The effects of predicting a person's behavior on subsequent trait judgments. *Journal of Experimental Social Psychology, 20,* 29-46.

Zedeck, S. & Cascio, W. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology*, *67*, 752-758.

Footnotes

[1] Since the manipulation of appraisal purpose was unsuccessful in the present study, one could argue that the inclusion of purpose in the analyses used to test the assumptions of processing invariance would possibly attenuate the effects of job familiarity on subjects' recall and ratings of ratee performance. In order to test this possibility, additional analyses were conducted on both the process and outcome measures without using purpose as a factor in the design. Results of these analyses were consistent with the analyses performed when purpose was included in the design. No other significant effects for job familiarity were obtained.

Table 1

Means and Standard Deviations for Manipulation Checks of Appraisal Purpose and Job Familiarity by Appraisal Purpose and Rater Population

	Rater Population			
	Students		Carpenters	
DV	Memory (n=20)	Impression (n=20)	Memory (n=20)	Impression (n=20)
Q#1	3.85 (0.88)	3.50 (0.95)	3.90 (0.72)	3.70 (0.80)
Q#2	4.00 (0.79)	4.15 (0.59)	4.30 (0.86)	4.50 (0.51)
Q#3	3.55 (0.83)	3.40 (0.94)	4.30 (0.66)	4.45 (0.83)
Q#4	7.55 (2.44)	6.45 (1.79)	18.50 (1.85)	18.10 (1.99)
Q#5	7.40 (3.87)	6.65 (2.06)	5.20 (2.61)	4.20 (1.51)
Q#6	3.40 (1.14)	3.20 (1.32)	2.05 (1.19)	1.65 (0.99)

Note: Q#1=extent to which the observational purpose was clear
Q#2=extent to which subjects tried to form an impression
Q#3=extent to which subjects tried to memorize details
Q#4=results of carpentry test (max=22)
Q#5=number of aspects of good teaching
Q#6=number of critical apsects of good teaching

Note: Numbers in parentheses denote standard deviations

Table 2

Means and Standard Deviations for Free Recall Measures by Appraisal Purpose and Rater Population Summed Across Performance Level

	Rater Population			
	Students		Carpenters	
Variable	Memory (n=20)	Impression (n=20)	Memory (n=20)	Impression (n=20)
TEACHING				
Judgments	9.75 (4.51)	11.70 (4.33)	7.95 (4.01)	6.70 (2.64)
Behaviors	7.70 (3.28)	4.35 (2.66)	4.90 (2.99)	3.55 (2.11)
Behaviors w/ Judg'ts	0.70 (0.98)	0.65 (0.88)	0.35 (0.81)	0.30 (0.66)
Total Recall	18.95 (3.95)	16.70 (4.32)	13.20 (3.66)	10.55 (3.41)
CARPENTRY				
Judgments	6.55 (5.02)	7.70 (5.92)	7.85 (4.39)	5.35 (4.00)
Behaviors	9.75 (5.76)	8.30 (5.52)	4.80 (3.40)	5.55 (4.55)
Behaviors w/Judg'ts	0.60 (0.88)	0.70 (0.80)	2.35 (1.95)	1.45 (1.67)
Total Recall	16.90 (3.91)	16.70 (4.51)	15.00 (4.23)	12.35 (3.20)

Note: Numbers in parentheses denote standard deviations

Table 3

MANOVA Table for Number of Behaviors and Number of Judgments Recalled by Rater Population and Recall for Subjects Receiving a Memory-Set

Source	Sums of Squares	DF	Mean Square	F-Ratio
BETWEEN-SUBJECTS				
Rater Population (R)	170.16	1	170.16	17.35**
Residual	372.69	38	9.81	
WITHIN-SUBJECTS				
Recall (C)	61.26	1	61.26	2.13
R X C	131.41	1	131.41	4.57*
Residual	1092.59	38	28.75	
Occup (O)	4.56	1	4.56	0.72
R X O	2.26	1	2.26	0.36
Residual	239.44	38	6.30	
C X O	68.91	1	68.91	2.46
R X C X O	68.91	1	68.91	2.46
Residual	1062.44	38	27.96	

Note: ** $p<.001$; * $p<.05$

Table 4

ANOVA Table for the Number of Judgments Recalled for the Target Occupation of Teaching

Source	Sums of Squares	DF	Mean Square	F-Ratio
Rater Population (R)	231.20	1	231.20	14.80**
Purpose (P)	2.45	1	2.45	0.16
R X P	51.20	1	51.20	3.28
Residual	1187.10	76	15.62	

Note: ** p<.001

Table 5

MANOVA Tables for the Number of Judgments Recalled for each Rater Population Across Target Occupations

Source	Sums of Squares	DF	Mean Square	F-Ratio
		STUDENTS		
BETWEEN-SUBJECTS				
Purpose (P)	48.05	1	48.05	0.28
Residual	1485.50	38	39.09	
WITHIN-SUBJECTS				
Occupation (O)	259.20	1	259.20	24.53**
P X O	3.20	1	3.20	0.30
Residual	401.60	38	10.57	
		CARPENTERS		
BETWEEN-SUBJECTS				
Purpose (P)	70.31	1	70.31	4.81*
Residual	555.07	38	14.61	
WITHIN-SUBJECTS				
Occupation (O)	10.51	1	10.51	0.71
P X O	7.81	1	7.81	0.53
Residual	561.17	38	14.77	

Note: ** $p < .001$
Note: * $p < .05$

Table 6

MANOVA Tables for the Number of Behaviors Recalled for each Rater Population Across Target Occupations

Source	Sums of Squares	DF	Mean Square	F-Ratio
		STUDENTS		
BETWEEN-SUBJECTS				
Purpose (P)	115.20	1	115.20	5.81*
Residual	753.75	38	19.84	
WITHIN-SUBJECTS				
Occupation (O)	180.00	1	180.00	8.63**
P X O	18.05	1	18.05	0.86
Residual	792.95	38	20.87	
		CARPENTERS		
BETWEEN-SUBJECTS				
Purpose (P)	1.80	1	1.80	0.13
Residual	514.00	38	13.53	
WITHIN-SUBJECTS				
Occupation (O)	18.05	1	18.05	1.94
P X O	22.05	1	22.05	2.37
Residual	352.90	38	9.29	

Note: ** $p < .01$
Note: * $p < .05$

Table 7

Means and Standard Deviations for Four ARC' Measures of Seriation by Appraisal Purpose and Rater Population

	Rater Population			
	Students		Carpenters	
DV	Memory (n=20)	Impression (n=20)	Memory (n=20)	Impression (n=20)
ARC' AVG	0.16 (0.36)	0.46 (0.49)	0.28 (0.44)	0.24 (0.46)
ARC' GOOD	0.12 (0.26)	0.49 (0.47)	0.15 (0.37)	0.37 (0.48)
ARC' POOR	0.33 (0.47)	0.13 (0.13)	0.37 (0.48)	0.25 (0.44)
ARC' TOT	0.21 (0.27)	0.36 (0.25)	0.27 (0.31)	0.29 (0.29)

Note: ARC'-AVG=ARC' score for average carpentry performance
Note: ARC'-GOOD=ARC' score for good carpentry performance
Note: ARC'-POOR=ARC' score for poor carpentry performance
Note: ARC'-TOT=composite ARC' score averaged over all 3 ratees

Note: Numbers in parentheses denote standard deviations

Table 8

MANOVA Tables for Proportion of Judgments Recalled for each Rater Population Across Target Occupations

Source	Sums of Squares	DF	Mean Square	F-Ratio
		STUDENTS		
BETWEEN-SUBJECTS				
Purpose (P)	0.25	1	0.25	0.08
Residual	2.82	38	0.07	
WITHIN-SUBJECTS				
Occupation (O)	0.71	1	0.71	15.03**
P X O	0.06	1	0.06	1.27
Residual	1.80	38	0.05	
		CARPENTERS		
BETWEEN-SUBJECTS				
Purpose (P)	0.00	1	0.00	0.06
Residual	2.27	38	0.06	
WITHIN-SUBJECTS				
Occupation (O)	0.42	1	0.42	7.54*
P X O	0.09	1	0.09	1.61
Residual	2.11	38	0.06	

Note: ** $p < .001$
Note: * $p < .01$

Table 9

MANOVA Tables for Proportion of Behaviors Recalled for each Rater Population Across Target Occupations

Source	Sums of Squares	DF	Mean Square	F-Ratio
		STUDENTS		
BETWEEN-SUBJECTS				
Purpose (P)	0.27	1	0.27	3.82
Residual	2.65	38	0.07	
WITHIN-SUBJECTS				
Occupation (O)	0.70	1	0.70	13.12**
P X O	0.06	1	0.06	1.11
Residual	2.03	38	0.05	
		CARPENTERS		
BETWEEN-SUBJECTS				
Purpose (P)	0.02	1	0.02	0.34
Residual	2.66	38	0.07	
WITHIN-SUBJECTS				
Occupation (O)	0.01	1	0.01	0.27
P X O	0.16	1	0.16	2.91
Residual	2.13	38	0.06	

Note: ** *p*<.001

Table 10

Means and Standard Deviations for Behavioral Ratings of Average Teaching Performance by Appraisal Purpose and Rater Population

	Rater Population			
	Students		Carpenters	
Variable	Memory (n=20)	Impression (n=20)	Memory (n=20)	Impression (n=20)
Q#1	5.20 (1.01)	5.50 (1.28)	5.40 (1.19)	4.80 (1.36)
Q#2	4.95 (1.43)	5.30 (1.42)	5.20 (1.82)	4.60 (1.79)
Q#3	4.55 (1.67)	5.20 (1.80)	4.75 (1.77)	4.00 (1.38)
Q#4	3.20 (1.36)	4.10 (1.71)	3.35 (1.69)	3.40 (1.57)
Q#5	3.50 (1.24)	3.85 (1.23)	3.80 (1.70)	3.85 (1.42)
Q#6	2.90 (1.02)	3.20 (1.11)	3.45 (1.54)	3.40 (1.14)
Q#7	3.15 (1.35)	3.55 (1.57)	3.65 (1.35)	3.20 (1.58)
Q#8	2.75 (1.25)	3.50 (1.53)	2.90 (1.68)	3.05 (1.82)
Q#9	4.60 (1.60)	5.20 (1.44)	4.95 (1.57)	4.75 (1.37)
Q#10	2.55 (1.50)	2.95 (1.82)	2.05 (1.10)	2.80 (1.28)
Q#11	4.75 (1.80)	4.95 (1.73)	5.00 (1.41)	4.15 (1.60)
Q#12	2.65 (1.23)	3.25 (1.45)	2.95 (1.50)	2.95 (1.64)
Q#13	4.45 (1.70)	4.25 (1.41)	4.60 (1.76)	5.10 (1.48)

Note: Numbers in parentheses denote standard deviations

Table 10 (cont.)

Note: Q#1=Uses hand or body movements to stress points
Q#2=Reads from notes
Q#3=Speaks in a monotone voice for a long period
Q#4=Frowns or makes distracting faces
Q#5=Makes eye contact with the audience
Q#6=Allows his voice to trail off or fade
Q#7=Hesitates, says "um" or "ah"
Q#8=Loses place or train of thought
Q#9=Changes his tone of voice or loudness to stress points
Q#10=Makes distracting hand or body movments
Q#11=Speaks more slowly when presenting complicated material
Q#12=Mumbles or fails to speak clearly
Q#13=Presents specific concrete examples

note: Questions 1, 5, 9, 11, and 13 were reverse-coded.

Table 11

Means and Standard Deviations for Behavioral Ratings of Good Teaching Performance by Appraisal Purpose and Rater Population

	Rater Population			
	Students		Carpenters	
Variable	Memory (n=20)	Impression (n=20)	Memory (n=20)	Impression (n=20)
Q#1	4.50 (1.47)	4.10 (1.41)	3.70 (1.84)	4.05 (1.67)
Q#2	4.10 (1.71)	3.95 (1.05)	4.30 (1.63)	4.10 (1.59)
Q#3	3.00 (1.65)	2.30 (1.26)	2.35 (1.35)	3.15 (2.01)
Q#4	2.20 (1.15)	2.40 (1.43)	2.15 (1.14)	2.45 (1.47)
Q#5	3.25 (1.21)	2.50 (0.69)	2.55 (0.83)	2.60 (0.94)
Q#6	2.35 (1.00)	2.30 (1.08)	2.05 (1.19)	2.40 (1.47)
Q#7	2.45 (1.19)	2.55 (1.05)	2.25 (1.41)	2.40 (1.27)
Q#8	2.15 (1.04)	1.90 (0.64)	1.70 (0.73)	2.30 (1.49)
Q#9	3.30 (1.72)	2.55 (1.61)	3.05 (1.32)	4.15 (1.95)
Q#10	2.40 (1.39)	2.25 (1.33)	2.60 (1.67)	1.75 (0.85)
Q#11	3.25 (1.77)	2.95 (1.32)	2.70 (1.63)	3.65 (2.01)
Q#12	1.85 (0.88)	1.70 (0.87)	1.60 (0.94)	1.80 (1.20)
Q#13	3.40 (1.54)	3.60 (1.79)	3.55 (2.01)	3.40 (1.90)

Note: Numbers in parentheses denote standard deviations

Table 11 (cont.)

Note: Q#1=Uses hand or body movements to stress points
Q#2=Reads from notes
Q#3=Speaks in a monotone voice for a long period
Q#4=Frowns or makes distracting faces
Q#5=Makes eye contact with the audience
Q#6=Allows his voice to trail off or fade
Q#7=Hesitates, says "um" or "ah"
Q#8=Loses place or train of thought
Q#9=Changes his tone of voice or loudness to stress points
Q#10=Makes distracting hand or body movments
Q#11=Speaks more slowly when presenting complicated material
Q#12=Mumbles or fails to speak clearly
Q#13=Presents specific concrete examples

Note: Questions 1, 5, 9, 11, and 13 were reverse-coded.

Table 12

Means and Standard Deviations for Behavioral Ratings of Poor Teaching Performance by Appraisal Purpose and Rater Population

	Rater Population			
	Students		Carpenters	
Variable	Memory (n=20)	Impression (n=20)	Memory (n=20)	Impression (n=20)
Q#1	6.30 (0.73)	6.55 (0.51)	6.35 (0.93)	5.75 (1.74)
Q#2	6.65 (0.49)	6.45 (0.95)	6.50 (0.76)	6.05 (1.43)
Q#3	6.05 (1.19)	6.05 (1.54)	6.35 (1.35)	6.05 (1.47)
Q#4	4.30 (1.81)	4.30 (1.49)	4.25 (1.94)	4.15 (1.60)
Q#5	5.90 (0.91)	5.90 (0.55)	5.70 (1.17)	5.30 (1.46)
Q#6	5.15 (1.18)	5.50 (1.05)	5.40 (1.43)	4.95 (1.67)
Q#7	4.35 (1.95)	4.05 (1.61)	4.35 (1.46)	4.00 (1.49)
Q#8	4.25 (1.25)	3.95 (1.36)	4.15 (1.50)	4.35 (1.69)
Q#9	6.25 (0.72)	6.30 (0.66)	5.95 (1.19)	5.65 (1.35)
Q#10	4.90 (2.08)	3.85 (2.16)	3.25 (1.62)	4.35 (1.95)
Q#11	6.15 (1.27)	6.05 (1.19)	5.40 (1.96)	5.05 (1.76)
Q#12	6.15 (0.93)	6.05 (1.32)	6.10 (1.30)	5.50 (1.40)
Q#13	5.30 (1.34)	5.10 (1.33)	5.25 (2.07)	5.10 (1.65)

Note: Numbers in parentheses denote standard deviations

Table 12 (cont.)

Note: Q#1=Uses hand or body movements to stress points
Q#2=Reads from notes
Q#3=Speaks in a monotone voice for a long period
Q#4=Frowns or makes distracting faces
Q#5=Makes eye contact with the audience
Q#6=Allows his voice to trail off or fade
Q#7=Hesitates, says "um" or "ah"
Q#8=Loses place or train of thought
Q#9=Changes his tone of voice or loudness to stress points
Q#10=Makes distracting hand or body movments
Q#11=Speaks more slowly when presenting complicated material
Q#12=Mumbles or fails to speak clearly
Q#13=Presents specific concrete examples

Note: Questions 1, 5, 9, 11, and 13 were reverse-coded.

Table 13

Means and Standard Deviations for Behavioral Ratings of Average Carpentry Performance by Appraisal Purpose and Rater Population

	Rater Population			
	Students		Carpenters	
Variable	Memory (n=20)	Impression (n=20)	Memory (n=20)	Impression (n=20)
Q#1	1.75 (1.29)	1.90 (1.65)	2.10 (1.92)	2.90 (1.92)
Q#2	1.05 (0.22)	1.05 (0.22)	1.35 (1.35)	1.90 (1.62)
Q#3	1.95 (1.82)	2.65 (1.98)	1.70 (1.30)	2.85 (1.95)
Q#4	3.50 (2.37)	2.95 (2.16)	4.45 (2.57)	4.05 (2.11)
Q#5	1.05 (0.22)	1.60 (1.35)	1.10 (0.31)	1.55 (1.47)
Q#6	3.65 (2.64)	3.35 (2.39)	2.25 (2.36)	3.80 (2.76)
Q#7	3.25 (2.34)	2.35 (1.93)	5.45 (2.14)	5.10 (2.47)
Q#8	1.80 (1.11)	1.80 (1.01)	2.10 (1.41)	2.75 (1.48)
Q#9	1.70 (1.30)	1.70 (1.34)	1.95 (1.47)	2.80 (1.91)
Q#10	1.95 (1.23)	1.85 (0.67)	3.25 (1.68)	3.10 (1.94)
Q#11	1.15 (0.37)	1.45 (0.76)	1.40 (0.94)	2.10 (1.68)
Q#12	2.25 (1.83)	2.75 (1.97)	4.20 (2.38)	4.80 (2.19)
Q#13	1.95 (1.40)	1.90 (1.12)	3.50 (2.48)	3.25 (2.10)

Note: Numbers in parentheses denote standard deviations

Table 13 (cont.)

Note: Q#1=Stains wood against the grain
Q#2=Uses a punch to set nails
Q#3=Rounds edges while sanding
Q#4=Sands wood with the grain
Q#5=Bends nails while hammering
Q#6=Uses a square to measure wood
Q#7=Saws wood in a straight line
Q#8=Drips or streaks stain
Q#9=Stains with the grain
Q#10=Bends the saw while sawing
Q#11=Splits the wood while hammering
Q#12=Uses full strokes while sanding
Q#13=Aligns edges of wood

Note: Questions 2, 4, 6, 7, 9, 12, and 13 were reverse-coded.

Table 14

Means and Standard Deviations for Behavioral Ratings of Good Carpentry Performance by Appraisal Purpose and Rater Population

	Rater Population			
	Students		Carpenters	
	Memory	Impression	Memory	Impression
Variable	(n=20)	(n=20)	(n=20)	(n=20)
Q#1	2.75 (1.77)	3.10 (1.92)	2.75 (2.15)	3.35 (1.79)
Q#2	2.10 (2.17)	1.55 (1.28)	1.95 (1.97)	2.45 (2.44)
Q#3	2.40 (2.14)	3.10 (1.89)	2.25 (1.92)	2.35 (1.76)
Q#4	2.90 (2.20)	2.40 (1.14)	2.05 (1.32)	2.30 (1.42)
Q#5	4.00 (1.78)	4.50 (1.43)	3.50 (1.99)	3.30 (1.78)
Q#6	3.20 (2.57)	3.80 (2.19)	2.90 (2.40)	2.85 (2.37)
Q#7	2.70 (1.84)	3.50 (1.64)	2.45 (1.93)	2.25 (2.00)
Q#8	2.90 (1.65)	2.85 (1.14)	2.60 (1.68)	3.30 (1.63)
Q#9	2.45 (1.28)	2.60 (1.23)	1.90 (1.37)	2.80 (1.61)
Q#10	3.40 (1.96)	3.40 (1.85)	3.30 (1.95)	3.25 (1.65)
Q#11	2.25 (1.71)	2.70 (1.38)	2.10 (1.71)	2.20 (1.64)
Q#12	3.00 (1.65)	3.45 (1.64)	4.15 (2.37)	3.45 (2.06)
Q#13	3.80 (2.40)	3.55 (1.93)	3.05 (2.37)	2.85 (2.01)

Note: Numbers in parentheses denote standard deviations

Table 14 (cont.)

Note: Q#1=Stains wood against the grain
Q#2=Uses a punch to set nails
Q#3=Rounds edges while sanding
Q#4=Sands wood with the grain
Q#5=Bends nails while hammering
Q#6=Uses a square to measure wood
Q#7=Saws wood in a straight line
Q#8=Drips or streaks stain
Q#9=Stains with the grain
Q#10=Bends the saw while sawing
Q#11=Splits the wood while hammering
Q#12=Uses full strokes while sanding
Q#13=Aligns edges of wood

Note: Questions 2, 4, 6, 7, 9, 12, and 13 were reverse-coded.

Table 15

Means and Standard Deviations for Behavioral Ratings of Poor Carpentry Performance by Appraisal Purpose and Rater Population

	Rater Population			
	Students		Carpenters	
Variable	Memory (n=20)	Impression (n=20)	Memory (n=20)	Impression (n=20)
Q#1	4.15 (1.18)	4.60 (1.39)	4.30 (1.90)	3.35 (1.79)
Q#2	6.10 (2.20)	5.60 (2.35)	6.10 (2.92)	4.50 (2.71)
Q#3	5.20 (2.19)	4.90 (2.10)	5.65 (2.18)	4.85 (2.06)
Q#4	2.20 (1.64)	2.80 (1.51)	2.45 (1.96)	2.45 (2.04)
Q#5	1.95 (1.97)	2.20 (1.20)	2.25 (1.62)	2.70 (2.03)
Q#6	4.45 (2.59)	3.95 (2.26)	1.95 (1.99)	2.70 (2.45)
Q#7	2.25 (1.16)	2.35 (0.88)	2.25 (1.83)	2.55 (2.14)
Q#8	2.75 (1.65)	2.95 (1.00)	3.80 (1.96)	3.15 (1.46)
Q#9	3.95 (1.43)	4.35 (1.60)	4.10 (1.62)	3.50 (2.07)
Q#10	3.25 (1.77)	2.75 (1.02)	3.00 (2.03)	2.90 (1.86)
Q#11	2.10 (1.94)	2.05 (0.76)	2.80 (2.14)	2.55 (1.91)
Q#12	2.70 (1.66)	2.50 (1.24)	4.25 (2.27)	3.00 (1.65)
Q#13	3.15 (2.21)	2.70 (1.34)	2.75 (2.36)	2.70 (1.87)

Note: Numbers in parentheses denote standard deviations

Table 15 (cont.)

Note: Q#1=Stains wood against the grain
Q#2=Uses a punch to set nails
Q#3=Rounds edges while sanding
Q#4=Sands wood with the grain
Q#5=Bends nails while hammering
Q#6=Uses a square to measure wood
Q#7=Saws wood in a straight line
Q#8=Drips or streaks stain
Q#9=Stains with the grain
Q#10=Bends the saw while sawing
Q#11=Splits the wood while hammering
Q#12=Uses full strokes while sanding
Q#13=Aligns edges of wood

Note: Questions 2, 4, 6, 7, 9, 12, and 13 were reverse-coded.

Table 16

Means and Standard Deviations for Graphic Ratings for each Level of Teaching Performance by Appraisal Purpose and Rater Population

	Rater Population			
	Students		Carpenters	
	Memory	Impression	Memory	Impression
Variable	(n=20)	(n=20)	(n=20)	(n=20)
Q#1A	3.05 (0.83)	2.95 (0.76)	3.15 (0.88)	3.10 (0.85)
Q#2A	2.75 (0.91)	2.80 (0.89)	3.20 (0.95)	3.15 (1.04)
Q#3A	3.00 (0.80)	2.65 (0.93)	3.20 (0.95)	3.30 (0.87)
Q#4A	3.00 (1.08)	2.90 (1.02)	2.90 (1.02)	2.90 (1.17)
Q#5A	2.75 (0.79)	2.55 (0.83)	3.10 (0.79)	3.20 (0.89)
Q#6G	4.00 (1.17)	4.10 (0.91)	3.75 (0.91)	4.00 (1.03)
Q#7G	3.65 (0.81)	3.75 (0.72)	3.95 (0.89)	4.15 (0.99)
Q#8G	4.00 (1.03)	4.10 (1.02)	4.10 (0.97)	4.45 (0.69)
Q#9G	4.00 (0.65)	3.80 (0.89)	4.15 (0.67)	4.30 (0.73)
Q#10G	3.75 (0.72)	3.85 (0.88)	4.10 (0.72)	4.25 (0.97)
Q#11P	1.30 (0.47)	1.85 (0.88)	1.95 (0.61)	2.05 (0.83)
Q#12P	2.40 (1.10)	2.20 (0.95)	2.60 (0.94)	2.70 (0.98)
Q#13P	1.30 (0.47)	1.30 (0.57)	1.30 (0.47)	1.95 (0.95)
Q#14P	2.70 (1.30)	2.40 (1.00)	2.70 (1.13)	2.55 (1.00)
Q#15P	1.75 (0.55)	1.80 (0.62)	1.95 (0.61)	2.10 (0.64)

Note: Q#1A,G,P=Teacher's organization and clarity
Q#2A,G,P=Teacher's responsivness to questions
Q#3A,G,P=Teacher's speaking ability
Q#4A,G,P=Teacher's poise and composure
Q#5A,G,P=Teacher's overall performance rating

Note: Numbers in parentheses denote standard deviations

Table 17

Means and Standard Deviations for Graphic Ratings for each Level of Carpentry Performance by Appraisal Purpose and Rater Population

	Rater Population			
	Students		Carpenters	
Variable	Memory (n=20)	Impression (n=20)	Memory (n=20)	Impression (n=20)
Q#1A	3.75 (1.21)	4.10 (0.97)	2.70 (1.17)	2.05 (1.05)
Q#2A	4.65 (0.59)	4.70 (0.47)	3.65 (1.04)	3.40 (0.83)
Q#3A	4.25 (0.85)	4.35 (0.67)	3.60 (0.99)	3.00 (0.97)
Q#4A	3.55 (1.28)	3.60 (1.27)	2.35 (1.31)	2.30 (1.38)
Q#5A	4.10 (0.97)	4.25 (0.72)	3.10 (0.85)	2.80 (0.62)
Q#6G	2.85 (1.39)	2.45 (1.10)	2.85 (1.14)	3.25 (1.02)
Q#7G	2.70 (0.92)	2.55 (1.15)	2.80 (1.01)	2.95 (1.05)
Q#8G	3.50 (1.24)	3.20 (0.77)	2.95 (0.99)	3.05 (0.91)
Q#9G	3.55 (1.05)	3.25 (0.79)	3.10 (0.97)	3.30 (0.73)
Q#10G	3.05 (0.95)	2.95 (0.89)	2.95 (0.89)	3.15 (0.59)
Q#11P	3.60 (0.88)	3.65 (0.67)	3.10 (1.17)	3.40 (1.14)
Q#12P	2.80 (0.95)	2.70 (0.87)	2.40 (0.88)	2.80 (1.01)
Q#13P	2.90 (1.12)	2.75 (0.79)	2.15 (0.88)	2.40 (0.88)
Q#14P	3.15 (1.35)	3.05 (1.19)	2.10 (0.91)	2.15 (0.88)
Q#15P	3.05 (0.95)	3.00 (0.65)	2.35 (0.75)	2.70 (0.73)

Note: Q#1A,G,P=How well the Carpenter sawed
Q#2A,G,P=How well the Carpenter hammered
Q#3A,G,P=How well the Carpenter stained
Q#4A,G,P=How well the Carpenter sanded
Q#5A,G,P=Carpenter's overall performance rating

Note: Numbers in parentheses denote standard deviations

Table 18

Means and Standard Deviations for Summed Behavioral Ratings for both Teaching and Carpentry Occupations by Level of Performance, Purpose, and Rater Population

	Rater Population			
	Students		Carpenters	
Variable	Memory (n=20)	Impression (n=20)	Memory (n=20)	Impression (n=20)
TEACHING				
Average	49.20 (10.00)	54.80 (10.41)	52.05 (10.22)	50.05 (0.47)
Good	38.20 (9.94)	35.05 (10.51)	34.55 (9.89)	38.20 (9.40)
Poor	71.70 (8.30)	70.10 (6.32)	69.00 (7.46)	66.25 (9.49)
CARPENTRY				
Average	27.00 (8.43)	27.30 (8.32)	34.80 (7.93)	40.95 (10.68)
Good	37.85 (13.61)	40.50 (9.92)	34.95 (12.40)	36.70 (8.86)
Poor	44.20 (13.03)	43.70 (7.70)	45.65 (9.97)	40.90 (10.36)

Note: Higher summed scores denote poorer ratee behavior

Note: Means are based on summed responses across each 13-item behavioral rating scale

Note: Numbers in parentheses denote standard deviations

Table 19

Means and Standard Deviations for Summed Graphic Ratings for both Teaching and Carpentry Occupations by Level of Performance, Purpose, and Rater Population

	Rater Population			
	Students		Carpenters	
Variable	Memory (n=20)	Impression (n=20)	Memory (n=20)	Impression (n=20)
TEACHING				
Average	14.55 (3.12)	13.85 (3.54)	15.55 (3.66)	15.65 (3.39)
Good	19.40 (3.10)	19.60 (3.90)	20.05 (3.36)	21.15 (3.07)
Poor	9.45 (2.78)	9.55 (2.72)	10.50 (2.76)	11.35 (2.58)
CARPENTRY				
Average	20.30 (3.70)	21.00 (3.33)	15.35 (4.04)	13.55 (3.28)
Good	15.70 (4.26)	14.40 (4.50)	14.65 (3.84)	15.70 (2.87)
Poor	15.50 (4.22)	15.15 (2.37)	12.10 (3.13)	13.45 (3.42)

Note: Higher summed scores denote better ratee performance

Note: Means are based on summed responses across each 5-item graphic rating scale

Note: Numbers in parentheses denote standard deviations

Table 20

Means and Standard Deviations for True Scores by Performance Level for Behavioral and Graphic Ratings for the Occupation of Teaching

Variable	Average	Performance Good	Poor
BEHAVIORS			
1. Hand Movements	6.10 (0.74)	5.20 (1.55)	6.80 (0.42)
2. Reads from Notes	5.00 (0.94)	4.50 (1.65)	6.10 (0.57)
3. Monotone Voice	3.30 (1.34)	1.70 (0.95)	6.20 (0.79)
4. Frowns	2.00 (0.82)	1.90 (0.74)	4.10 (1.37)
5. Makes Eye Contact	3.00 (0.82)	2.70 (0.82)	5.40 (0.84)
6. Voice Fades	2.50 (1.51)	2.00 (1.49)	5.10 (1.19)
7. Hesitates	3.20 (1.14)	2.10 (0.74)	4.40 (0.70)
8. Loses Place	2.10 (0.74)	1.80 (0.79)	3.80 (1.03)
9. Changes Tone	3.60 (1.43)	2.10 (0.99)	6.00 (0.47)
10. Distracting Body Movements	1.60 (0.84)	3.00 (1.49)	3.60 (1.35)
11. Speaks Slowly	5.10 (1.10)	3.80 (1.13)	6.20 (0.63)
12. Mumbles	1.80 (0.63)	1.30 (0.68)	5.70 (0.82)
13. Concrete Examples	4.40 (1.17)	4.70 (1.06)	5.00 (1.15)

Note: Numbers in parentheses denote standard deviations

Note: Behaviors 1, 5, 9, 11, and 13 were reverse-coded

Table 20 (cont.)

Means and Standard Deviations for True Scores by Performance Level for Behavioral and Graphic Ratings for the Occupation of Teaching

		Performance	
Variable	Average	Good	Poor
JUDGMENTS			
1. Organization/ Clarity	3.50 (0.97)	4.20 (1.03)	1.70 (0.48)
2. Response to Q's	2.70 (0.68)	3.90 (0.57)	2.60 (0.70)
3. Speaking Ability	3.50 (0.53)	4.50 (0.71)	1.30 (0.48)
4. Grasps Material	2.90 (0.74)	4.50 (0.71)	2.00 (0.68)
5. Overall Rating	3.30 (0.68)	4.40 (0.53)	1.40 (0.53)

Note: Numbers in parentheses denote standard deviations

Table 21

Means and Standard Deviations for True Scores by Performance Level for Behavioral and Graphic Ratings for the Occupation of Carpentry

Variable	Performance Average	Good	Poor
BEHAVIORS			
1. Stains Ag/ Grain	1.20 (0.45)	4.40 (1.13)	4.40 (2.07)
2. Uses Punch	1.00 (0.00)	1.80 (1.30)	4.40 (2.61)
3. Rounds Edges	1.60 (0.89)	1.20 (0.46)	5.60 (1.14)
4. Sands w/ Grain	4.60 (1.82)	1.00 (0.00)	2.00 (1.73)
5. Bends Nails	1.00 (0.00)	5.00 (0.71)	1.60 (0.89)
6. Uses Square	1.20 (0.45)	1.40 (0.89)	1.40 (0.55)
7. Saws Straight	5.80 (2.17)	1.60 (0.55)	2.00 (0.71)
8. Streaks Stain	1.60 (0.89)	1.00 (0.00)	1.80 (1.09)
9. Stains w/ Grain	1.20 (0.45)	1.20 (0.45)	4.00 (1.22)
10. Bends Saw	1.60 (0.55)	2.80 (2.39)	2.80 (1.92)
11. Splits Wood	1.20 (0.45)	1.40 (0.55)	2.00 (1.22)
12. Full Strokes/Sand	4.60 (2.19)	3.20 (1.79)	4.20 (2.05)
13. Aligns Edges/Wood	2.60 (1.52)	3.00 (2.00)	2.60 (0.89)
JUDGMENTS			
1. Saw	1.60 (0.89)	3.80 (0.45)	3.40 (0.89)
2. Hammer	4.60 (0.55)	2.00 (0.71)	1.60 (0.55)
3. Stain	4.00 (0.99)	3.20 (1.48)	2.00 (0.71)
4. Sand	1.40 (0.55)	3.80 (0.84)	1.80 (0.84)
5. Overall Rating	2.60 (0.55)	3.40 (0.55)	1.80 (0.45)

Note: Numbers in parentheses denote standard deviations

Note: Behaviors 2, 4, 6, 7, 9, 12, and 13 were reverse-coded

Table 22

Means and Standard Deviations for Variance Components of Rating Accuracy for Behavioral and Graphic Ratings of Teaching and Carpentry by Purpose and Rater Population

	Rater Population			
	Students		Carpenters	
	Memory	Impression	Memory	Impression
Variable	(n=20)	(n=20)	(n=20)	(n=20)
TEACHING: BEHAVIORAL RATINGS				
Elevation	0.31 (0.25)	0.39 (0.33)	0.33 (0.31)	0.33 (0.36)
Diff Elev	0.60 (0.27)	0.65 (0.26)	0.58 (0.32)	0.51 (0.33)
Stereo Acc	0.82 (0.22)	0.81 (0.23)	0.89 (0.29)	0.95 (0.27)
Diff Acc	0.95 (0.19)	0.93 (0.24)	0.99 (0.32)	1.14 (0.40)
GRAPHIC RATINGS				
Elevation	0.27 (0.20)	0.38 (0.38)	0.32 (0.32)	0.20 (0.22)
Diff Elev	0.53 (0.24)	0.49 (0.24)	0.46 (0.27)	0.49 (0.31)
Stereo Acc	0.40 (0.16)	0.32 (0.14)	0.32 (0.13)	0.37 (0.21)
Diff Acc	0.54 (0.16)	0.50 (0.15)	0.48 (0.15)	0.54 (0.24)
CARPENTRY: BEHAVIORAL RATINGS				
Elevation	0.60 (0.34)	0.46 (0.29)	0.60 (0.41)	0.61 (0.34)
Diff Elev	0.60 (0.45)	0.59 (0.32)	0.50 (0.25)	0.62 (0.32)
Stereo Acc	1.19 (0.35)	1.11 (0.38)	1.14 (0.30)	1.21 (0.48)
Diff Acc	1.38 (0.32)	1.36 (0.22)	1.43 (0.26)	1.47 (0.35)
GRAPHIC RATINGS				
Elevation	0.70 (0.46)	0.64 (0.33)	0.47 (0.24)	0.33 (0.27)
Diff Elev	0.77 (0.41)	0.80 (0.39)	0.50 (0.28)	0.49 (0.25)
Stereo Acc	0.35 (0.16)	0.31 (0.16)	0.34 (0.09)	0.35 (0.10)
Diff Acc	1.46 (0.31)	1.39 (0.30)	1.23 (0.28)	1.24 (0.24)

Note: Numbers in parentheses denote standard deviations

Note: Smaller values denote higher levels of rating accuracy

Table 23

Means and Standard Deviations for Correlational Components of Rating Accuracy for Behavioral and Graphic Ratings of Teaching and Carpentry by Purpose and Rater Population

	Rater Population			
	Students		Carpenters	
Variable	Memory (n=20)	Impression (n=20)	Memory (n=20)	Impression (n=20)
TEACHING BEHAVIORAL RATINGS				
DECOR	0.89 (0.11)	0.87 (0.10)	0.85 (0.22)	0.86 (0.10)
SACOR	0.71 (0.13)	0.65 (0.29)	0.67 (0.22)	0.56 (0.28)
DACOR	0.42 (0.22)	0.35 (0.23)	0.28 (0.14)	0.37 (0.22)
GRAPHIC RATINGS				
DECOR	0.88 (0.14)	0.88 (0.16)	0.89 (0.16)	0.85 (0.26)
SACOR	-0.02 (0.45)	-0.08 (0.40)	-0.30 (0.35)	-0.10 (0.50)
DACOR	0.42 (0.22)	0.35 (0.23)	0.48 (0.19)	0.37 (0.22)
CARPENTRY BEHAVIORAL RATINGS				
DECOR	0.57 (0.52)	0.49 (0.58)	0.61 (0.51)	0.18 (0.79)
SACOR	0.04 (0.26)	0.11 (0.28)	0.34 (0.30)	0.21 (0.41)
DACOR	0.42 (0.22)	0.35 (0.23)	0.48 (0.19)	0.37 (0.22)
GRAPHIC RATINGS				
DECOR	0.05 (0.69)	0.08 (0.50)	0.44 (0.55)	0.35 (0.65)
SACOR	0.13 (0.63)	0.22 (0.55)	0.29 (0.44)	0.24 (0.36)
DACOR	0.37 (0.36)	0.36 (0.31)	0.45 (0.33)	0.47 (0.28)

Note: Larger values denote higher levels of rating accuracy

Table 24

MANOVA Table for Differential Accuracy on the Behavioral Rating Scales by Appraisal Purpose and Rater Population Across Target Occupations

Source		Sums of Squares	DF	Mean Square	F-Ratio
BETWEEN-SUBJECTS					
Rater Population	(R)	0.46	1	0.46	4.42*
Purpose	(P)	0.06	1	0.06	0.56
Residual		7.85	76	0.10	
WITHIN-SUBJECTS					
Occupation (O)		6.71	1	6.71	94.73**
R X O		0.02	1	0.02	0.32
P X O		0.03	1	0.03	0.38
R X P X O		0.03	1	0.03	0.44
Residual		5.38	76	0.07	

Note: ** $p<.001$
Note: * $p<.05$

Table 25

ANOVA Table for Elevation of Graphic Ratings for Carpentry Performance

Source	Sums of Squares	DF	Mean Square	F-Ratio
Rater Population (R)	1.41	1	1.41	12.57**
Purpose (P)	0.21	1	0.21	1.85
R X P	0.03	1	0.03	0.26
Residual	8.51	76	0.11	

Note: ** *p*<.001

Table 26

ANOVA Table for Differential Elevation of Graphic Ratings for Carpentry Performance

Source	Sums of Squares	DF	Mean Square	F-Ratio
Rater Population (R)	1.78	1	1.78	15.43**
Purpose (P)	0.00	1	0.00	0.02
R X P	0.01	1	0.01	0.07
Residual	8.76	76	0.12	

Note: ** $p<.001$

Table 27

ANOVA Table for Differential Accuracy of Graphic Ratings for Carpentry Performance

Source	Sums of Squares	DF	Mean Square	F-Ratio
Rater Population (R)	0.67	1	0.67	8.56*
Purpose (P)	0.02	1	0.02	0.20
R X P	0.03	1	0.03	0.37
Residual	6.10	76	0.08	

Note: * $p<.01$

Appendix A

Recalls of Ratee Behavior: Carpentry Performance

Focusing on carpenter "Mike", please LIST everything you can remember from the videotape. Please number each of your comments.

Recalls of Ratee Behavior: Teaching Performance

Focusing on teacher "Rusty", please LIST everything you can remember from the videotape. Please number each of your comments.

Appendix B

Behavioral Rating Scale: Carpentry Performance

Rate the frequency that carpenter "Mike" did the following behaviors. Use the following scale to rate the frequency of each behavior. Mark all of your answers in the spaces provided directly on this sheet.

1	2	3	4	5	6	7
Never	Almost Never	A Few Times	About Half of the Time	Often	Most of the Time	All of the Time

1. Stains wood against the grain ________

2. Uses a punch to set nails ________

3. Rounds edges while sanding ________

4. Sands wood with the grain ________

5. Bends nails while hammering ________

6. Uses a square to measure wood ________

7. Saws wood in a straight line ________

8. Drips or streaks stain ________

9. Stains with the grain ________

10. Bends the saw while sawing ________

11. Splits wood while hammering ________

12. Uses full strokes while sanding ________

13. Aligns edges of wood ________

Behavioral Rating Scale: Teaching Performance

Rate the frequency that teacher "Rusty" did the following behaviors. Use the following scale to rate the frequency of each behavior. Mark all of your answers in the spaces provided directly on this sheet.

1	2	3	4	5	6	7
Never	Almost Never	A Few Times	About Half of the Time	Often	Most of the Time	All of the Time

1. Uses hand or body movements to stress points ________

2. Reads from notes ________

3. Speaks in a monotone voice for a long period ________

4. Frowns or makes distracting faces ________

5. Makes eye contact with the audience ________

6. Allows his or her voice to trail off or fade ________

7. Hesitates, says "um" or "ah" ________

8. Loses place or train of thought ________

9. Changes his or her tone of voice or loudness to stress points ________

10. Makes distracting hand or body movements ________

11. Speaks more slowly when presenting complicated material ________

12. Mumbles or fails to speak clearly ________

13. Presents specific, concrete examples as part of his or her lecture ________

Appendix C

Graphic Rating Scale: Carpentry Performance

Rate the performance of carpenter "Mike" using the following scale. Mark all of your answers in the spaces provided directly in this sheet.

1	2	3	4	5
Very Bad	Poor	Average	Good	Very Good

1. How well did carpenter Mike saw? ______

2. How well did carpenter Mike hammer? ______

3. How well did carpenter Mike stain? ______

4. How well did carpenter Mike sand? ______

5. How would you rate carpenter Mike
on his OVERALL performance? ______

Graphic Rating Scale: Teaching Performance

Rate the performance of teacher "Rusty" using the following scale. Mark all of your answers in the spaces provided directly in this sheet.

1	2	3	4	5
Very Bad	Poor	Average	Good	Very Good

1. Teacher Rusty's organization and clarity ______

2. Teacher Rusty's responsiveness to questions ______

3. Teacher Rusty's speaking ability ______

4. Teacher Rusty's grasp of material ______

5. What is your overall rating of teacher Rusty's performance? ______

Appendix D

Manipulation Check Items

Please answer the following questions using the scales provided below. Please circle the number that best matches your response. Make sure you only choose 1 number, do not mark any responses in between numbers.

1. Using the following scale, report how CLEAR you perceived your given observational purpose.

1	2	3	4	5
Totally Unclear	A little Clear	Somewhat Clear	Clear	Very Clear

2. Using the following scale, how much did you TRY to form an IMPRESSION of the persons' performances.

1	2	3	4	5
None of the time	A few times	Some of the time	Most of the time	All of the time

3. Using the following scale, how much did you TRY to MEMORIZE details of the the persons' performances.

1	2	3	4	5
None of the time	A few times	Some of the time	Most of the time	All of the time

4. Using the following scale, how much did you either try to form an impression of the persons' performances or memorize details of their performances.

1	2	3	4	5
always formed impressions	mostly formed impressions	A little bit of both	mostly memorized	always memorized

Appendix E

Consent Forms

CONSENT FORM (Carpenters)

TO ALL RESEARCH PARTICIPANTS:

This experiment is being conducted under the supervision of Steven E. Walker; it is an investigation of the way in which persons view other persons' behavior in a work setting. We will be asking you to complete a job knowledge test and then view several different persons on videotape and then rate their performances. Below are listed some of the items of information that you should know when deciding to participate in this study.

(1) No psychological or physical harm is expected to result from your participation in the experiment.

(2) The experiment requires two hours of your time. You will be paid $40 for your participation, and your cooperation is greatly appreciated.

(3) Your agreement to participate should be voluntary and thus can be withdrawn at any time by you without penalty.

(4) All information gathered from your responses is intended for research purposes only. Therefore, your responses will remain completely confidential.

(5) This research project has been approved by the Human Subjects Committee of the Psychology department. Any questions you have can be answered by contacting one of the individuals listed below:

Dr. Helen Crawford (x6520): Human Subjects Committee
Dr. Neil Hauenstein: (office: 231-5716)
Steven E. Walker: (office: 231-8141)

(6) There is a copy of this consent form available if you should wish to retain a copy for your personal records, BUT PLEASE FILL OUT THE COPY THAT IS ATTACHED TO THE SURVEY FOR OUR RECORDS.

If you consent voluntarily and with an understanding of the conditions outlined above to participate in the experiment, please sign your name below. Thank you very much for your assistance.

signature______________________________

Name (please print): ___________________________

ID #:_________________________

CONSENT FORM (students)

TO ALL RESEARCH PARTICIPANTS:

This experiment is being conducted under the supervision of Steven E. Walker; it is an investigation of the way in which persons view the behavior of other persons in a work setting. We will be asking you to complete a job knowledge test and then view several different persons on videotape and then rate their performances. Below are listed some of the items of information that you should know when deciding to participate in this study.

(1) No psychological or physical harm is expected to result from your participation in the experiment.

(2) The experiment requires two hours of your time. You will earn two extra credit points towards your PSYC 2000 grade and your cooperation is greatly appreciated.

(3) Your agreement to participate should be voluntary and thus can be withdrawn at any time by you without penalty.

(4) All information gathered from your responses is intended for research purposes only. Therefore, your responses will remain completely confidential.

(5) This research project has been approved by the Human Subjects Committee of the Psychology department. Any questions you have can be answered by contacting one of the individuals listed below:

Dr. Helen Crawford (x6520): Human Subjects Committee
Dr. Neil Hauenstein: (office: 231-5716)
Steven E. Walker: (office: 231-8141)

(6) There is a copy of this consent form available if you should wish to retain a copy for your personal records, BUT PLEASE FILL OUT THE COPY THAT IS ATTACHED TO THE SURVEY FOR OUR RECORDS.

If you consent voluntarily and with an understanding of the conditions outlined above to participate in the experiment, please sign your name below. Thank you very much for your assistance.

signature______________________________

Name (please print): ___________________________

ID #:_________________________

Appendix F

Carpentry Test and Screening Items

CARPENTRY TEST

Instructions

DO NOT TURN THE PAGE UNTIL YOU ARE INSTRUCTED TO DO SO.

You will need a NUMBER 2 PENCIL to complete this test. If you do not have one, please ask the experimenter for a pencil.

Put your Social Security Number on your answer sheet.

This test contains 22 multiple-choice items covering various aspects of carpentry. Please answer each question to the best of your knowledge. Your score will be the total number of questions answered correctly. Please keep in mind that the test is timed. You will have 20 minutes to work.

Each question has four (4) possible answers: a, b, c, or d. On your OPSCAN sheet, darken the circle corresponding to the letter of the correct answer. Please make sure you COMPLETELY fill in the circle corresponding to the correct answer.

EXAMPLE: What tool is a "chuck-key" used with?

a. Saw
b. Hammer
c. Drill
d. Vise

The correct answer is "c". You would darken the circle of the letter "c" on your OPSCAN sheet.

EXAMPLE: a b c d

Your score on this test will be the total number of questions you answered correctly. If you are sure you know the answer to a question, circle the appropriate letter on the answer sheet. It will be to your advantage to answer questions in which you can eliminate one or more options. However, there is no penalty for guessing.

CARPENTRY TEST

1. What size level would you use to check if a wall is plumb?

   a. 9 inch
   b. 2 feet
   c. 4 feet
   d. 5 feet

2. What is a "catspaw" used for?

   a. extracting nails
   b. determining bevel cuts
   c. countersinking
   d. measuring angles

3. What type of saw is most often used for cutting base molding?

   a. miter
   b. saber
   c. circular
   d. hacksaw

4. How many inches are studs usually centered on?

   a. 14
   b. 16
   c. 18
   d. 20

5. What does a "2 X 4" actually measure?

   a. 1 1/4 X 3 1/4
   b. 1 1/2 X 3 1/2
   c. 1 3/4 X 3 3/4
   d. 2 X 4

6. What tool is used to determine lengths of common, hip, and valley rafters?

   a. framing square
   b. tape rule
   c. hand square
   d. bevel square

7. What size nail would you use to install shoe molding?

   a. 2d
   b. 4d
   c. 8d
   d. 10d

8. What tool has a rafter table stamped on its face?

   a. circular saw
   b. bevel square
   c. miter saw
   d. framing square

9. What grade of lumber is used for framing members such as studs, rafters, and joists?

   a. number 1 common
   b. number 2 common
   c. number 3 common
   d. number 4 common

10. What handsaw is used for general cutting?

   a. backsaw
   b. ripsaw
   c. hacksaw
   d. crosscut saw

11. What is another word for framing carpentry?

   a. finish carpentry
   b. trim carpentry
   c. rough carpentry
   d. cabinet building

12. What kind of interior trim goes around doors and windows?

   a. baseboard
   b. lattice
   c. molding
   d. casing

13. What is the exterior trim called at the point where roof projections and side walls meet?

    a. cornice
    b. crown molding
    c. rough sill
    d. wall apron

14. Which saw is best for cutting curved lines?

    a. coping saw
    b. circular saw
    c. hacksaw
    d. crosscut saw

15. What is the trim at the lower part of a window opening called?

    a. apron
    b. casing
    c. sill
    d. jamb

16. Which of the following are shims used for?

    a. bracing a wall
    b. padding out a door frame
    c. scabbing wood together
    d. anchoring a stud wall

17. Which of the following has a clutch?

    a. miter saw
    b. drill
    c. circular saw
    d. screwshooter

18. What kind of tool is best for planing door edges?

    a. block plane
    b. smooth plane
    c. jack plane
    d. jointer plane

19. What are tongue and groove boards typically used for?

    a. ceilings
    b. roofs
    c. floors
    d. walls

20. What is a "lintel?"

    a. a horizontal support used over doors and windows
    b. a saw used for making bevel cuts
    c. a drill used to bore into hardwoods
    d. a beam which runs across the center of a roof

21. What do you call pieces of wood placed between floor joists or wall studs for reinforcement?

    a. braces
    b. nailers
    c. bridges
    d. supports

22. How do you dull the tip of a nail?

    a. file the tip down
    b. snip the tip off
    c. hit the tip with a hammer
    d. bend the tip

Please list what you feel are the important aspects of good teaching performance. It may be helpful for you to include both good and poor examples of teaching performance. Please list and number your comments.

ANSWER KEY: Carpentry Test

1. c
2. a
3. a
4. b
5. b
6. a
7. b
8. d
9. b
10. d
11. c
12. d
13. a
14. a
15. a
16. b
17. d
18. d
19. c
20. a
21. c
22. c

CARPENTRY EXPERIENCE (carpenters)

Please answers the following questions to the best of your knowledge. Again, fill in the circles on your answer sheet that match with the letter on the questionnaire.

23. How many years of PROFESSIONAL carpentry experience do you have?

    a. none
    b. 1-3
    c. 4-7
    d. 8 or more

24. Are you, or have you ever been a member of a carpentry union? (please choose only one response)

    a. yes
    b. no

25. What type of carpentry are you specialized in? (Darken all responses that apply)
    a. rough
    b. finish
    c. cabinet building
    d. other

ACADEMIC EXPERIENCE (carpenters)

Please answer these questions to the best of your knowledge. Darken the circles on your answer sheet that match the letters on the questionnaire.

24. Please indicate the highest grade of schooling you have completed.

    a. 10th grade or less
    b. 11th grade
    c. 12th grade (graduated high school)
    d. 1 or more years of college

25. Only if you answered "d" to question 24, indicate what kind of COLLEGE education you have had. (Where appropriate, write down any additional comments directly on THIS sheet).

    a. night school or part-time student
    b. full-time student
    c. currently enrolled in college (indicate which college)
    d. graduated college with A.A., B.A., and/or B.S.

26. Have you ever been enrolled in any vocational schools? (If you answer "yes", describe your background directly on THIS sheet.)

    a. no
    b. yes

CARPENTRY EXPERIENCE (students)

Please answers the following questions to the best of your knowledge. Again, fill in the circles on your answer sheet that match with the letter on the questionnaire.

23. Have you ever had any PROFESSIONAL carpentry experience? If so, indicate how many years.

    a. none
    b. less than 1 year
    c. 2 years
    d. more than 2 years

24. Have you ever had a part-time or summer job in carpentry or construction? If so, indicate how many times you worked in this setting.

    a. never
    b. once
    c. twice
    d. more than twice

25. Do you have any IMMEDIATE relatives with any carpentry or construction experience? (Only relatives you have lived with).

    a. yes
    b. no

26. Do you have any SERIOUS hobbies involving woodworking or construction. If so, please describe this hobby(s) in the space provided below on THIS questionnaire.

    a. no
    b. yes
    [If you answered "b", explain in the space provided below]

27. Have you ever done any MANUAL labor that has been connected with carpentry or construction. If so, please describe your experience(s) in the space provided below on THIS questionnaire.

    a. no
    b. yes
    [If you answered "b", explain in the space provided below]

28. What is your academic major?

a. undeclared
b. psychology
c. engineering
d. architecture
e. liberal arts
f. computer science
g. mathematics
h. English
i. other

Appendix G

Computational Formula for the Adjusted Ratio of Clustering (ARC') Measure of Seriation.

$$ARC'=[O(ITR) - E(ITR)] / [MAX - E(ITR)]$$

where, *Max*=*M* - 1 - *R*
E(ITR)=2 (Max) / *N*

M= Number of items recalled on trial *t*
N= Number of items recalled on trial *t+1*
O(ITR)= Number of observed pairwise bidirectional repetitions
Max=Maximum number of bidirectional pairwise repetitions
E(ITR)= Expected number of bidirectional repetitions

and *t*= Input list
t ± *1*= Output list

From:

Murphy, M.D. & Puff, C.R. (1982). Free recall: Basic methodology and analyses. In C.R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 99-128). New York: Academic Press.

Appendix H

Computational Formulae for Variance and Correlational Measures of Rating Accuracy.

1. Algebraic Difference Score Formulae for the Four Components of Accuracy. For a rater who evaluates *n* ratees on *k* items or dimensions, scores on elevation (E), differential elevation (DE), stereotype accuracy (SA), and differential accuracy (DA) are given by the square root of the following terms:

$$E^2 = (\overline{x}.. - \overline{t}..)^2$$

$$DE^2 = 1/n\sum [\overline{x}i. - \overline{x}..) - (\overline{t}i. - \overline{t}..)]^2$$

$$SA^2 = 1/k\sum [\overline{x}.j - \overline{x}..) - (\overline{t}.j - \overline{t}..)]^2$$

$$DA^2 = 1/kn\sum\sum [(\overline{x}ij - \overline{x}i. - \overline{x}.j + \overline{x}..) -$$

$$(\overline{t}ij - \overline{t}i. - \overline{t}.j + \overline{t}..)]^2$$

where xij and tij = rating and true score for ratee *i* on item *j*; xi. and ti. = mean rating and mean true score for ratee *i*; x.j and t.j = mean rating and mean true score for item *j*; and x.. and t.. = mean rating and mean true score, over all ratees and items.

2. Accuracy Formulae in Terms of Variances and Correlations.

$$DE^2 = \sigma^2_{\overline{x}i.} + \sigma^2_{\overline{t}i.} - 2\,\sigma_{\overline{x}i.}\sigma_{\overline{t}i.}\, \underline{r}_{\overline{x}i.\overline{t}i.}$$

$$SA^2 = \sigma^2_{\overline{x}.j} + \sigma^2_{\overline{t}.j} - 2\,\sigma_{\overline{x}.j}\sigma_{\overline{t}.j}\, \underline{r}_{\overline{x}.j\overline{t}.j}\text{ , \underline{and}}$$

$$DA^2 = \sigma^{2\prime}_{xij} + \sigma^{2\prime}_{tij} - 2\sigma'_{xij}\,\sigma'_{tij}\ \underline{r}'_{xij\,tij}$$

where x'ij(t'ij) represents the rating and true score for ratee *i* on the *jth* item after subtracting $\overline{x}i.(\overline{t}i.)$ and $\overline{x}.j(\overline{t}.j)$ and adding $\overline{x}..(\overline{t}..)$. All other terms are defined above.

From:

Murphy, K.R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W.K. (1982). Relationship between observation accuracy and accuracy in evaluating performance. *Journal of Applied Psychology*, *67*, 320-325.

Becker, B.E. & Cardy, R.L. (1986). Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. *Journal of Applied Psychology*, *71*, 662-671.

The three page vita has been removed from the scanned document. Page 1 of 3

The three page vita has been removed from the scanned document. Page 2 of 3

The three page vita has been removed from the scanned document. Page 3 of 3